The Dissertation Committee for Huimin Zhao Certifies that this is the approved version of the following dissertation:

Toward a Comprehensive Hazard-Based Duration Framework to Accommodate Nonresponse in Panel Surveys

Committee:

Chandra R. Bhat, Supervisor

C. Michael Walton

Hani S. Mahmassani

Susan Handy

Lynne Stokes

Toward a Comprehensive Hazard-Based Duration Framework to Accommodate Nonresponse in Panel Surveys

by

Huimin Zhao, B.S., M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December, 2002

Dedication

To my grandmother

Acknowledgements

I am very grateful to many individuals who generously offered advice, encouragement, and friendship along my way to the completion of this dissertation. First of all, I would like to thank my advisor, Professor Chandra Bhat, for his expert guidance, mentoring, and research support throughout my doctoral study. Without his support this dissertation would not have been possible. Furthermore, I offer my sincere gratitude to the members of my dissertation committee. Many thanks go to Professor Hani S. Mahmassani, for his great classes which provided me with a fundamental understanding of transportation system analysis, and for his invaluable comments on the organization of this dissertation; to Professor C. Michael Walton, who offered an opportunity for me to see various aspects of engineering; to Professor Susan Handy, who helped me find useful references in the early stage of this research; and to Professor Lynn Stokes, who helped me see things from different prospective.

Thank you to my friends and fellow graduate students in the transportation program who made the 6th floor of ECJ Hall a pleasant place to work and learn. I would particularly like to thank Aruna Sivakumar, Nathan Nhan Huynh, and Jessica Guo for reading the early draft of this dissertation. On a more personal note, I want to thank my husband, Yong Zhao, for his infinite support, patience, and good temper. Last but certainly not least, my warmest thanks go to my

parents, for the way I was raised and countless philosophies of life I have been taught; and to my little sister, Hui-Qun, for sharing all the struggles and successes of my life so far.

Toward a Comprehensive Hazard-Based Duration Framework to Accommodate Nonresponse in Panel Surveys

Publication No._____

Huimin Zhao, Ph.D.

The University of Texas at Austin, 2002

Supervisor: Chandra R. Bhat

Many surveys suffer from low response rates and therefore carry a risk of nonresponse bias. The problem is more severe in panel surveys because sample units are subject to nonresponse repeatedly. This dissertation is concerned with nonresponse in longitudinal household travel surveys. It identifies the likely sources of nonresponse and investigates a model-based bias correction procedure for the subject of interest in the survey -- trip frequency.

Low response rates often lead to a sample representativeness problem and threaten the validity of the survey. A better understanding of the survey participation behavior can provide guidance for survey design to increase the response rates and to build an effective nonresponse bias correction procedure. It is generally believed that nonresponse is a combined result of social environment,

survey attributes, and characteristics of sample units. In addition, state dependence and the lagged impact of exogenous variables can not be ignored when considering repeated responses in panel surveys. The first stage of this work considers the repeated participation in panel surveys as a duration process and proposes a hazard-based duration model for the analysis. The model structure accommodates state dependence and the lagged effects in a straightforward manner. Various factors, especially the indicators of survey burden, are incorporated in the model for a comprehensive understanding of the survey participation decision. The empirical analysis based on the seven-wave Puget Sound Transportation Panel suggests that survey burden, in general, is negatively associated with the survey participation duration. The results also reflect an interactive impact of survey burden and time constraints on the survey participation. The second stage of this work further investigates the relationship between the survey participation and trip frequency. The model formulation incorporates observed and unobserved heterogeneity in the participation and travel decisions. It is found that trip frequency, especially for home-based non-work trips, is endogenously correlated with the survey participation decision and the ignorance of this endogenous correlation leads to a biased estimate for the trip frequency and survey participation duration.

Table of Contents

List of Tables xiii

List of Figures xv

Chapter 1 Introduction 1

1.1 Motivation 1

1.1.1 Need for Survey Guidance to Achieve Higher Response Rates 2

1.1.2 Importance of Consistent Parameter Estimation 2

1.1.3 Complex Behavioral Mechanism Observed in Panel Data 4

1.1.4 Common Unobserved Factors Associated with Survey Participation Decision and Travel Behavior 4

1.2 Objectives of the Dissertation 5

1.3 Overview of Panel Data 8

1.3.1 Advantages of Panel Data 9

1.3.2 Modeling Issues with Panel Data 13

1.4 Outline of the Dissertation 15

Chapter 2 Literature Review 18

2.1 Influence of Survey Design on Nonresponse 18

2.1.1 Understanding the Decision to Participate in a Survey 19

2.1.2 Influence of Survey Burden and Interviewers on Nonresponse 23

2.2 Post-survey Statistical Approach to Correct for Nonresponse Bias 28

2.2.1 Imputation and Weighting Methods 31

2.2.2 Empirical Studies 39

2.3 Attrition in Multi-wave Panel Data 43

2.4 Summary 46

Chapter 3 Research Approach 48
3.1 Modeling Discrete Variables in Panel Data 49
3.1.1 Discrete Choice Model and Choice Behavioral Theory 49
3.1.2 Choice Behavior in Panel Data 53
3.2 Travel Behavior and Nonresponse in Household Travel Survey 58
3.3 Summary 60

Chapter 4 Data Description 62
4.1 Introduction 62
4.2 Sample Evolution 65
4.3 Imputation for Item Non-response 72
4.3.1 Imputation for Household Demographics 72
4.3.2 Imputation for Trip Characteristics 75
4.4 Trends in Household Demographics 78
4.4.1 Household Location and Household Type 78
4.4.2 Household Income and Vehicle Ownership 85
4.4.3 Household Size and Workers 88
4.5 Trends in Travel Activities 91
4.6 Summary 93

Chapter 5 Analysis of Survey Participation Duration with Trip Frequencies as Exogenous Variables 101
5.1 Duration Model 101
5.1.1 Introduction 101
5.1.2 Distribution of Hazard Function 105
5.1.3 Proportional Hazard Duration Model 110
5.2 Modeling Survey Participation Duration with Trip Frequencies as Exogenous Variables 114
5.2.1 Model with No Heterogeneity 114
5.2.2 Model with Gamma Heterogeneity 119
5.3 Data Sets for Model Estimation and Validation 120

5.4 Empirical Results 122
5.3.1 Covariate Effects 123
5.3.2 Baseline Hazard Rate 136
5.5 Comparison of Hazard-Based Duration Model with Discrete Choice Model 139
5.5 Summary 144

Chapter 6 Capturing Observed and Unobserved Factors Associated with Survey Participation and Trip Frequency 146
6.1 Measurements of Survey Burden 147
6.2 Modeling Considerations 151
6.2.1 Correlations among Trip Frequency and Survey Participation Duration 151
6.2.2 Dynamic vs. Static Model for Panel Data 152
6.3 Model Structure 153
6.4 Model Estimation 158
6.4.1 Monte Carlo and Quasi-Monte Carlo Methods 158
6.4.2 Halton Sequence 161
6.4.3 Simulated Likelihood Function 166
6.5 Empirical Results 168
6.5.1 Survey Participation 168
6.5.2 Home-Based Work Trips 172
6.5.3 Home-Based Non-Work Trips 173
6.6 Summary 175

Chapter 7 Conclusions 184
7.1 Contributions 184
7.2 Summary of Findings 185
7.3 Applications, Recommendations, and Future Research 188
7.3.1 Applications 188
7.3.2 Recommendations for Survey Design 189

7.3.3 Recommendations for Initial Nonresponse Study and Future Research 190

References 192

Vita..... 205

List of Tables

Table 4-1: Frequency of dropout households 69
Table 4-2: Survey participation of households with travel data missing 77
Table 4-3: Households by residence location and wave of the survey 79
Table 4-4a: Households by residence location and participation duration 80
Table 4-4b: Participation duration for relocated households 82
Table 4-5: Households by life cycle type and wave 83
Table 4-6: Households by life cycle type and survey participation duration 85
Table 4-7: Households by income category 86
Table 4-8: Households by vehicle ownership 87
Table 4-9: Households by household size 89
Table 4-10: Number of split households by wave 90
Table 4-11: Households by the number of workers 91
Table 4-12: Households by sampling group and survey participation duration ... 92
Table 5-1: Households' survey participation duration in calibration and validation sets 121
Table 5-2: Summary statistics for the hazard models 123
Table 5-3: Covariates effect of household demographics 126
Table 5-4: Covariate effect of survey burden, sampling group, and others 127
Table 5-5: Estimated coefficients for the logit model 141
Table 5-6: Comparison of the sign of estimated parameters 142
Table 5-7: Goodness-of-fit measures in validation data 143
Table 6-1: Survey participation duration in joint model system 176

Table 6-1: Survey participation duration in joint model system (cont.) 177
Table 6-2: Model for home-based work trips..... 178
Table 6-3: Model for home-based non-work trips 179
Table 6-4: Random-coefficient ordered response probit model for HBW trips (without accommodating selectivity bias)..... 180
Table 6-5: Standard ordered response probit model for HBW trips (without accommodating selectivity bias) 181
Table 6-6: Random-coefficient ordered response probit model for HBNW trips (without accommodating selectivity bias) 182
Table 6-7: Ordered response probit model for HBNW trips (without accommodating selectivity bias) 183

List of Figures

Figure 1-1: An example of biased estimation due to non-response 3
Figure 3-1: Elements in decision-making process (cited from McFadden,2000) . 51
Figure 4-1: Sample stratifications of the PSTP 68
Figure 4-2: Household dropout percentage 70
Figure 4-3: Households survey participation duration 71
Figure 4-4: Trip frequency for home-based work trips 95
Figure 4-5: Trip frequency for home-based non-work trips 96
Figure 4-6: Trip frequency for non-home-based trips 97
Figure 4-7: Trip frequency for driving-along trips 98
Figure 4-8: Trip frequency for carpool/vanpool trips 99
Figure 4-9: Trip rates by travel mode (transit trips) 100
Figure 5-2: Layout of the non-parametric baseline hazard rates 116
Figure 5-2: Baseline hazard rate (no heterogeneity) 138
Figure 5-3: Baseline hazard rate (Gamma heterogeneity) 138

Chapter 1 Introduction

Data constitute an essential component of transportation modeling, and sample surveys are considered as a fact-finding instrument for transportation modelers. However, non-response affects the quality and cost of surveys and even risks the validity of inferences from a survey. Households and/or individuals may not respond to surveys for different reasons. This dissertation evaluates this non-response phenomenon in multi-wave household travel surveys, as well as examines the endogenous correlation between non-response and trip frequency. The dissertation aims to better understand households' survey participation decision, and contributes to the development of survey strategy guidelines for non-response reduction. The dissertation also proposes an innovative approach to control for nonrandom non-response in model estimation from panel data.

1.1 MOTIVATION

This dissertation is motivated by four factors:

- The need for guidance to achieve higher response rates through effective survey design and administration;
- The importance of consistent parameter estimations using survey data;
- The complex behavioral mechanism of survey participation and travel decisions observed in panel surveys;

- The hypotheses of common unobserved factors affecting both survey participation and travel behavior.

1.1.1 Need for Survey Guidance to Achieve Higher Response Rates

The effective identification of the likely sources of non-response can reduce survey costs and ensure the timeliness and quality of surveys. It often occurs that a sample survey fails to obtain responses from some sample units. Some of the non-respondents may refuse to provide information, while some others may not be available for response. Survey quality may fall when sample size decreases due to non-response, and survey costs may increase to recruit more sample units. It is generally believed that the characteristics of survey design and data collection procedure, interacted with sample units' demographic features, determine response rates in surveys. The interactions, the *ad hoc* aspect of survey features, and the high cost of conducting controlled experiments make it difficult to pinpoint the causes of non-response. There is also a lack of substantive quantitative studies on how survey design and administration efforts can effectively improve the response rate, even though the topic has been long studied by survey researchers, statisticians, and economists.

1.1.2 Importance of Consistent Parameter Estimation

In general, it is not appropriate to estimate a model based on respondents alone, since the non-respondents may be different from the respondents in certain systematic ways. Unbiased estimations can not be obtained without a non-

response bias correction procedure. Figure 1-1 illustrates a simple example of how non-random non-response can lead to a biased estimation. Suppose non-respondents in a household travel survey happen to be low-income households. The task is to examine the relationship between trip rates and household income, and assume that an ordinary least squares linear regression model is estimated. Part (a) shows the linear regression based on the respondents. The demonstrated relationship between trip rates and income is valid only when non-response is completely random, which is often not the case in practice. In practice, non-response often occurs systematically. When the survey non-response is endogenously correlated with trip rates, the true regression line, which could be totally different as shown in part (b), may be obtained though an appropriate non-response bias correction procedure.

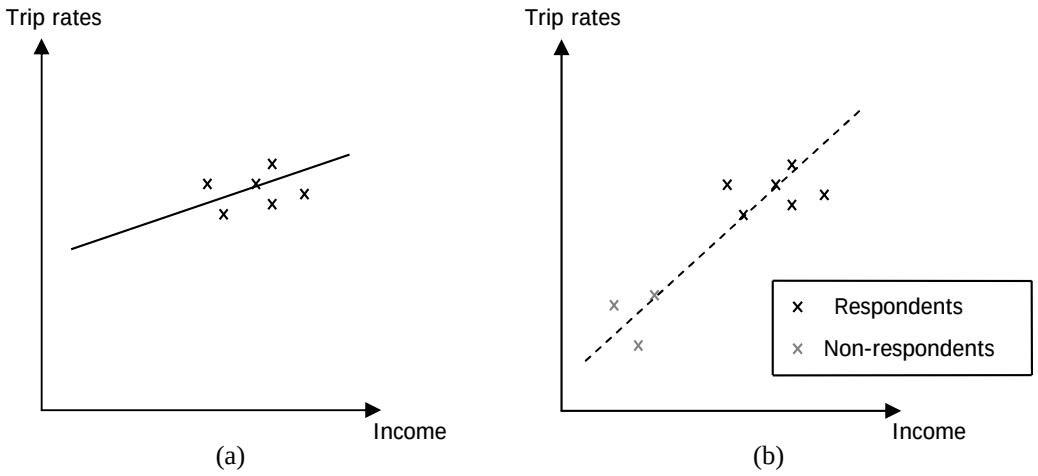

Figure 1-1: An example of biased estimation due to non-response

1.1.3 Complex Behavioral Mechanism Observed in Panel Data

The use of panel surveys and data in transportation study imposes substantial challenges on the non-response bias correction procedure. Collecting panel data is far more difficult than collecting cross-sectional data in terms of survey design and data management. Similarly, non-response in panel surveys merits careful consideration since the survey participants are subject to non-response, also referred to as panel attrition, repeatedly in the following waves. At the same time, the advantages of repeated observations and the comparability across panel waves from a behavioral perspective have not yet been fully explored in survey non-response studies. A comprehensive analysis of various impacts on non-response, including both survey features and sample unit characteristics, seems essential to build up standards for non-response measurement and to provide insights for effective survey planning.

1.1.4 Common Unobserved Factors Associated with Survey Participation Decision and Travel Behavior

Non-response may be associated with the survey subjects in one way or another. A household's decision to participate in a travel survey may explicitly depend on the number of travel activities it made during the survey period. Or, the participation decision may be associated with travel activities through some observed household characteristics. Furthermore, the correlation may be affected by some common unobserved factors and is likely to be endogenous. For instance, Kitamura and Bovy (1987) found strong correlation between the residuals of the trip frequency model and attrition probability based on two-wave

Dutch National Mobility Panel data. Consequently, it is necessary to test these behavioral hypotheses through a flexible model framework.

1.2 OBJECTIVES OF THE DISSERTATION

Driven by the motivations discussed in the previous section, the objectives of this dissertation are threefold. The first objective is to consistently model panel attrition for multiple waves. One common purpose of attrition modeling is to develop a weighting mechanism to correct for the non-response bias. The weighting system relies on the correct specification of the attrition model. Many earlier nonresponse studies have been taken in the context of cross-sectional data or one wave of panel data. Although the same methodology can be applied to the sequential panel waves repeatedly, lagged effects and state dependence across panel waves call for a more realistic model that considers multiple waves simultaneously. For example, a sample unit's experience in the first wave may play an important role in participation in the following waves. Moreover, the decision of staying in the panel survey may depend on the cumulative effect of the entire past experience. Modeling panel attrition separately from wave to wave cannot fully accommodate these effects. The lack of empirical studies on multi-wave attrition modeling may be due to the complexity of the model structure and computational burden of model estimation. In this dissertation, we specify a hazard-based duration model for the attrition process. The model structure has the advantage of incorporating state dependency without involving substantial computational burden.

The second objective of this study is to provide insights into the effectiveness of survey strategies on survey non-response. Panel attrition models can be used not only for the weighting system, but also for understanding the survey response behavior which can help to improve the efficiency of future survey design and therefore increase the survey response rate. Hensher (1987) and Horowitz (1997) suggested that survey operations should be undertaken in a way that maximizes response rate. It is essential to keep the attrition rate low while data are being collected. Approaches such as reminder calls, tracing efforts, and questionnaire updating can be used to achieve higher response rates. These efforts can be more efficiently conducted with the non-response segment successfully targeted. The post-survey quantitative analysis of attrition behavior offers an opportunity to identify the group which is more likely to not respond. Additional data collection efforts may be carried out for this targeted group. Therefore, identifying this group can make survey operations more cost efficient. The descriptive statistics approach has been widely adopted in survey-effectiveness studies in the past, but it is often not able to distinguish the survey features that can effectively improve the response rate from the others. This is partly because non-response is a joint result of all survey features, and partly because making an appropriate comparison of the non-response rates across surveys with different subjects is very difficult. The work presented in this dissertation adopts a quantitative method to overcome these difficulties.

Moreover, the approach employed in this study captures the longitudinal effect along panel waves and identifies the panel fatigue point. In panel surveys,

one common operational decision to make in panel surveys is how many waves of data to collect. In practice, funding surely is the most important factor. In addition, the number of waves is also determined by the purpose of the panel survey. For panels which are collected to evaluate a policy change, for instance, two or three waves of data may be collected before and after the policy change. For panels which are collected for general purposes, panel fatigue is another factor that needs to be considered in survey design. Panel fatigue occurs when the response rate declines after several waves because of boredom and related phenomena (Raimond & Hensher, 1997). Clearly, efforts to reduce the attrition rate beyond the fatigue point are likely to be far more costly than those pursued before the onset of fatigue. Therefore, understanding panel fatigue can also help to conduct surveys more efficiently. Simply comparing attrition rates in different waves may not provide convincing conclusions on panel fatigue because refreshment samples enter the survey in different waves. In the current research, we consider the panel survey participation as a duration process. In the duration model, the baseline hazard rate reflects the impact of the current spell of participation on the attrition probability and can be considered as a measure of panel fatigue. More details on the hazard model are discussed in Chapter 4.

The third objective of this dissertation is to examine the impact of survey burden on attrition and the correlation among panel attrition and trip frequency. It is a common belief that survey burden has a negative impact on survey participation. In household travel surveys, eligible household members are asked to record their travel activities during the survey period. Trip frequency can be

considered as an indicator of survey burden and may influence households' decision to participate in the survey. Therefore, the selectivity bias caused by panel attrition needs to be accommodated to obtain consistent estimates of the trip frequency model. Because of the endogenous nature of the correlation between the attrition process and trip frequency, a sequential modeling system for attrition and trip frequency will lead to biased and inefficient estimates. We formulate a model system that considers attrition and trip frequency simultaneously for multiple waves. This model system aims to demonstrate the impact of survey burden on attrition as well as to consistently analyze the trends in trip frequencies over time.

The work described in this dissertation is based on panel data. The next section provides an overview of panel data. The overview covers two sections: the advantages of panel data and the modeling issues associated with the use of panel data.

1.3 OVERVIEW OF PANEL DATA

Panel surveys record repeated measurements on the same set of households or individuals at different points in time. Panel data initially were of interest to economists in the mid-1960s. Two of the most prominent panel data in economics are the national longitudinal survey of labor market, started in 1966, and the University of Michigan's panel study of income dynamics, started in 1968. Panel data have provided researchers an opportunity to build and test more realistic behavior models that cannot be identified using cross-sectional and time

series data. Meanwhile, new modeling issues such as correction for heterogeneity bias due to individual and/or time characteristics have been raised with the use of panel data.

It was not until the 1980s that panel data were used for travel behavior analysis. The Dutch National Mobility Panel (DNMP) was first collected during a three-week period in March, 1984, and a second wave in September, 1984. The objectives of the DNMP survey were to evaluate changes in mobility over time and to assess the impact of transit fare increases on mobility. In the DNMP survey, a stratified sampling method was used to select the panel households with household lifecycle, income, and availability of various public transit modes as controlling factors. The data comprise seven-day diaries filled out by household members who were 12 years of age or older. As one of the earliest established transportation panels, the DNMP has provided a powerful research tool for understanding the changes in travel behavior. In the United States, the Puget Sound Transportation Panel (PSTP) is the first general-purpose transportation panel survey. The first wave of the panel was collected in 1989 and the survey continues to be administrated periodically.

1.3.1 Advantages of Panel Data

The most obvious advantage of transportation panel data over cross-sectional data is that panel data collect not only household demographic characteristics and travel information, but also changes in these variables over time. Panel data can be used to study the relationships between past, present, and

future experiences. Knowing such relationships helps analysts to understand the dynamics of travel behavior, and how an individual's travel choices may evolve in the future. It is difficult to make inferences about behavioral dynamics using cross-sectional data, for which the measurements are recorded at a single point in time. When using cross-sectional data, the inferences are drawn on the basis of the differences across individuals. When using panel data, the inferences are not only based on the differences across individuals, but also on the behavior differences in the same individual over time.

The models estimated from the cross-sectional data are referred to as static models in the literature (Tourangeau *et al.*, 1997). Static models assume that the effects of independent variables are instantaneous and that the relationship among variables is invariant over time. Consider a case study to model the impact of income on household vehicle ownership. The assumptions imply the following: a). a change in household income immediately results in a change in household vehicle ownership; b). the relationship between income and number of vehicles remains the same over time. Recent studies have challenged the validity of these assumptions (Goodwin, 1997; Golob, 1990). Golob (1990) studied the dynamics of household travel time expenditures and car ownership decisions using a pooled wave-pair sample drawn from four waves of the DNMP data. His study showed that there were important dynamic effects between household car ownership and household characteristics. For example, there was a lag on the effects of income and household drivers on vehicle ownership; current vehicle ownership depended on the household income in the past year. This result indicated that the changes

do not happen instantaneously. Because of the inertial effect, the household vehicle ownership not only depends on the present household characteristics, but also on the household characteristics in the previous year. With household demographic information at a single point in time, cross-sectional data cannot capture this inertial effect.

Panel data have two substantial advantages over cross-sectional data. First, dynamic behavior cannot be observed in the cross-sectional survey. These behavioral phenomena include responses to variable changes that are related to the time scale, such as the effect of household characteristics in the previous year on current vehicle ownership found in the study by Golob (1990). For this reason, studies of information acquisition, learning experience, and behavioral change can be accurately examined only using panel data (Kitamura, 1989). Second, even though some behavior can be observed in the cross-sectional data, the observations may be misleading if not correctly explained. In other words, the models established by cross-sectional data do not necessarily capture the actual relationships among variables. This is another advantage of panel data -- its ability to reveal true causality.

A classic example of using panel data to reveal the true causality is the effect of unionism on economic behavior (Freeman and Medoff, 1981; Hsiao, 1986). It is widely believed that unions and the bargaining process have a fundamental impact on the employment relationship, such as salary and work environment. This effect can be measured by including a dummy variable representing the presence of unions in the wage equation. The coefficient can be

explained as the union's power for higher wage. On the other hand, if one believes that firms hire higher-quality workers to react to the higher wage forced by the union (assuming the quality of the workers are not observable), the coefficient of the dummy variable simply indicates the worker quality. The cross-sectional data usually cannot distinguish the difference between these two alternative causality hypotheses. However, if panel data are used for the study, these two hypotheses can be tested by analyzing the wage difference for a worker who moved from a non-union firm to a union firm and thus reveal the true causality of the higher wage. If unions do have a positive impact on wage, the worker's wage will rise as he moves from a non-union firm to a union firm. In this case, the true causality of the wage raise is captured by maintaining the unobserved worker quality as a constant.

In summary, panel data can be used to develop more accurate and efficient econometric models. Kitamura (1990) pointed out that the models can uncover more precise behavioral changes when the unobserved attributes are well controlled using panel data. Hsiao (1986) indicated that panel data improve the efficiency of econometric estimates because the data provide a large number of observation points and thus increase the degrees of freedom among observed variables. In addition, panels can also be used to assess the impact of policy changes, such as the impact of a gasoline tax on household vehicle ownership. Because gasoline tax is usually a constant within a state over a certain period of time, the evaluation model cannot be identified without repeated measurements that trace the change in gasoline tax.

1.3.2 Modeling Issues with Panel Data

Although panel data have advantages for analyzing the impact of changes over time, several modeling issues need to be considered in utilizing panel data. Hsiao (1986) suggested that two types of econometric considerations need to be accommodated when using panel data. One is unobserved heterogeneity and the other is selectivity.

Unobserved heterogeneity arises when the conditional parametric probability distributions of dependent variable *y*, given independent variables *x's*, are not identical across individuals. While unobserved heterogeneity may also exist in cross-sectional data, it is more likely to appear in panel data because survey respondents repeatedly participate in surveys. Consequently, unobserved individual-specific effects lead to varying behavior among individuals and a positive correlation along observations on the same individual. In a standard model structure, the unobserved effect is captured by a random variable that is identically independently distributed (IID). This IID assumption cannot reflect the positive correlations among panel survey participants. There are several methods that may be used to capture unobserved heterogeneity in panel data. These methods are discussed in detail in Chapter 6.

The other issue that needs to be accommodated while using panel data is selectivity. Selectivity bias occurs when samples are not randomly drawn from the population and therefore cannot represent the population. Representativeness is an essential feature of survey data. In some cases, samples are randomly

selected as realizations of the entire population. However, the initial contacts are not always successful. Some sample households refuse to respond to the survey and then the sample does not fully represent the population. Thus, selectivity bias is often associated with data collection problems. In other cases, the population is divided into strata. The sample units are randomly drawn from each stratum instead of the whole population and selectivity bias arises. The stratified sampling approach offers a better representation for rare events and constitutes considerable cost savings for the survey. In travel surveys, for instance, the stratified sampling method is often used to exaggerate the fraction of the transit user group in the sample to efficiently model travel mode choice.

Selectivity bias exists in both cross-sectional data and panel data. Nevertheless, compared to cross-sectional surveys, panel surveys confront more challenges in collecting accurate data and representing the population. Some panels can last for years. In the national longitudinal survey of labor market, for example, four sets of sub-samples had been interviewed periodically for fifteen years. During the survey period, population and sample demographics change. For regional panel surveys, the characteristics of the population can vary dramatically as households move into or out of the region. However, the demographic changes in the sample data are gradual rather than abrupt if no refreshment samples are recruited. Therefore, the original panel sample has a potential problem in representing the population over time.

A major source of selectivity bias in panel surveys is panel attrition. For example, the attrition rate in the DNMP was at a disturbingly high level of 32%

between the first and second wave (Kitamura & Bovy, 1987). If panel attrition occurs randomly, the remaining sample can still represent the population and panel attrition would not lead to invalid inferences as long as the sample size is maintained at a reasonable size. However, many studies have suggested that panel attrition does not occur randomly. In the DNMP, for instance, attrition rates were higher among low-income, smaller, and less-educated households. Therefore, this group of households was under-represented in the second wave. Applying standard modeling procedure without considering attrition bias will lead to inconsistent estimates and invalid conclusions.

This dissertation presents an in-depth investigation into accommodating both heterogeneity and selectivity bias using panel data. The proposed modeling framework is then applied to explore the relationship between two key subjects of longitudinal household travel survey, panel attrition and trip frequency.

1.4 OUTLINE OF THE DISSERTATION

The remainder of this dissertation is organized as follows. Chapter 2 reviews the literature on survey non-response and panel attrition. Three topics are covered in this chapter. Theories on survey participation behavior in survey and psychology literature are reviewed first, followed by the post-survey quantitative methods used to correct for non-response bias. The section of post-survey quantitative methods presents the structural models formulated by econometricians and empirical studies undertaken by transportation analysts.

Finally, the review efforts focus on multi-wave panel attrition analyses that are mainly conducted in economics.

Based on the review of literature, Chapter 3 describes the research approach adopted in this dissertation. The survey participation decision is often modeled as a discrete variable in earlier studies. The formulation of realistic behavioral models for discrete variables in panel data and the challenges of model estimation are discussed at first in this chapter. The application challenges lead this research to consider non-response patterns through the use of hazard-based duration structure. Then, the potential hypotheses of the relationship between survey participation and the mobility indicators are presented, which leads to a joint model formulation of survey participation and mobility decisions to accommodate the potential correlation between them.

Chapter 4 presents descriptive statistics of household demographics and travel characteristics in the seven-wave PSTP data used for empirical analysis in the current research. The data collection procedure for the PSTP is briefly introduced in the first section. The data description then focuses on three aspects. The sample evolution throughout the seven waves is illustrated first, followed by a depiction of household demographic changes, and finally a description of changes in travel characteristics.

The first section of Chapter 5 provides a brief review of duration models. A hazard-based duration model for survey participation duration is also described in this chapter. Three issues are considered in the hazard-based duration model structure. These issues are the baseline hazard function, the effect of time-varying

covariates, and heterogeneity across individuals. A non-parametric form is adopted for the baseline hazard function. Meanwhile, the effect of time-varying covariates is implemented in the model structure. In addition, unobserved heterogeneity is captured by a random term with gamma distribution. The empirical results of the duration model with trip frequency as an exogenous variable are also presented in this chapter.

In Chapter 6, we formulate a model system that estimates the duration of survey participation and trip frequencies for multiple waves simultaneously. A common disturbance term is introduced to accommodate the endogenous correlation of the duration process and trip frequency model. The introduction of multi-level random effects in the model increases the computational burden for model estimation. To overcome the computational difficulties and improve the efficiency of the estimation procedure, we adopt quasi-Monte Carlo simulation techniques for model estimation.

The last chapter, Chapter 7, summarizes the findings of this study and provides a discussion on future research topics.

Chapter 2 Literature Review

It is both important to minimize non-response rate by survey design and necessary to reduce non-response bias by post-survey adjustment. The literature review covers these two main subjects. First, we discuss the influence of survey design and interviewers on non-response. Researchers in sociology and psychology have been trying to interpret who, where, and why non-response occurs. Survey experts have also tried to understand how different survey methods and the role of interviewers may have some impact on survey participation. The review summarizes the previous studies on non-response and the effectiveness of survey design on non-response. Second, we evaluate the post-survey statistical approach to adjust for non-response bias. Because the objective of this study is to correct for non-response bias in panel data, more extensive reviews are undertaken in the second section. Two classic non-response bias correction methods (imputation and weighting methods) are reviewed, followed by empirical studies in transportation fields. In the last section of this chapter, modeling techniques to adjust for multi-wave panel attrition are visited. Some empirical analyses, mainly undertaken in economics field, are also reviewed.

2.1 Influence of Survey Design on Non-response

One obvious consequence of non-response in a survey is that sample size is smaller than was originally planned. The smaller sample size is generally due

to two types of non-response: *unit* and *item* non-response. Sometimes a mail questionnaire is not returned, or a telephone interview is turned down. Consequently, none of the variables are measured for a household or an individual. This type of non-response is called *unit non-response*. The other type of non-response is *item non-response*. In this case, some of the questions for a unit are answered except for those that are usually sensitive, e.g., questions for income information.

2.1.1 Understanding the Decision to Participate in a Survey

Survey research is an effective method for describing the characteristics of a large population. Moreover, the value of the inferences drawn from a sample heavily depends on the representativeness of the sample. Indeed, the inferences may not be reliable when the response rate is too low. Steegh (1981) and Bogen (1996) have shown that the non-response rate is increasing, which makes it more challenging for survey designers to reduce non-response. Understanding why sample units choose to participate in a survey will help survey designers develop data collection procedures to achieve higher response rate. Many factors can influence response rate, such as household characteristics, and the interviewers' persuasive ability. Generally response rate is a result of the interaction of these factors. It is challenging to systematically differentiate the interactions and discover the cause of non-response.

The causes of non-response may be classified into two categories. One is the factors that survey operators can control while they generally do not have full control on the other. Sometimes a questionnaire that is too long or an interviewer

who has less experience may result in a high non-response rate. In this case, the response rate can be improved by a better survey design and training for the interviewers. In other words, the survey researchers can control these factors that cause non-response. In other cases, the non-response behavior may be due to the characteristics of survey participants and the social environment surrounding them. Survey researchers have little control over these factors. Generally, alternative post-survey analysis methods are applied to correct for the non-response bias resulting from such factors.

Atrostic *et al.* (1999) examined trends in response rates in six large continuing federal household surveys based on three projects undertaken by the Interagency Household Survey Nonresponse Group (IHSNG). The authors analyzed the initial non-response rates by categorizing the observed non-response rate into five classes. These five classes are refusals, no one at home, temporary absence, language problems, and an "all other reasons" category. Because many of these household surveys were panel surveys, the paper also presents a comparison of the response rates among each panel wave, as well as the response and non-response rates among the different surveys.

Through descriptive statistics, the study showed that refusals were the major source at the initial interview and the "no one at home" category became a more important component for non-response thereafter. Nonetheless, the non-response rate in each category varied widely from survey to survey. In addition, the comparison of response rates for different panel waves indicated that, throughout the duration of the panel, the refusal rate changed. However, the

change in refusal rates did not follow any systematic pattern after the first interview. Sometimes it decreased and sometimes it increased. The study also compared survey-specific non-response rates in terms of sample unit (person vs. household), and data collection mode (paper-and-pencil vs. telephone). Similarly, the response rates varied widely. By comparing the non-response rate in different panel surveys, Astrostic *et al.* (1999) highlighted the need to develop systematic non-response measures that are well defined and can address the various causes of non-response. The paper also recommends that additional research be conducted to identify the most important design features that may affect non-response rate through the duration of the panel.

Given the complex nature of survey non-response, the interagency efforts by Astrostic *et al.* (1999) aim to provide a broader and more systematic review than any one single agency could manage. However, the subject or topic of the survey might play a role in survey non-response, while the descriptive statistics approach adopted by the authors cannot effectively assess the separate impact of the survey subject as many changes occur simultaneously in the surveys. The subject matter may partially account for the inconsistent patterns of non-response rates across surveys in the IHSNG study. For instance, the IHSNG (1999) found higher response rates in the National Health Interview Survey (NHIS) and the National Crime Victimization Survey (NCVS), perhaps because people care about their health and consider participating in the NCVS as a means to reduce crime. A similar phenomenon is also observed in the Labor Force Survey in Ireland where employment is a very important topic (de Heer and Moritz, 1997).

Following the same logic, it seems plausible to explicitly inform the potential survey participants in household travel surveys that the information they provide in surveys would help ease traffic jams and build more livable communities.

The subject matter, often reflecting the interest value and the personal relevance of the survey, can be viewed as part of the psychological concept of compliance with request. In social psychology, research focusing on the relevant issue of survey participation can be divided into three areas: compliance with requests, helping tendencies, and opinion change (Groves *et al.*, 1992). Among these three areas, compliance with requests has more direct connection to the decision to participate in a survey because people often make a decision of performing an activity on the basis of the attractiveness (or unattractiveness) of the activity itself. Many survey operational strategies actually follow the compliance principles. For instance, the money incentive, often used in surveys to obtain higher level of survey cooperation, follows the reciprocity rule because people feel obligated to respond when receiving money incentives. Research on helping tendencies views survey request as a helping request under non-emergency situation. It has been found that three emotional states, anger, happiness, and sadness, are connected to helping decisions. Understanding what people's emotional states may be can help interviewers to react better in the first contact. If the household appears angry, it would be more successful if the interviewer retreats and returns later. Research on opinion change found that when survey topic is of high personal relevance, potential participants would base their opinion on intrinsic features of the topic itself; when the topic is less

personal important, they will change their opinion on the basis of its extrinsic features, such as interpersonal and societal factors.

As psychologists try to systematically categorize the rules of performing an activity, some argue that the decision of survey participation is often made in a heuristic manner. According to a theory of survey participation, a person's decision to participate in a survey generally occurs during the first moments of interaction with an interviewer or with the survey content (Couper and Groves, 1996). The respondent's mood may vary during the day, and consequently, the timing of the first telephone contact also contributes to this heuristic process.

The combination of the systematic and heuristic process can probably describe the nature of many decisions that we make in everyday life, given the rationality of the human being, imperfect information, and uncertainty in the future. The challenge arises in quantifying the relative contributions of systematic and heuristic impacts in decision making. Some actions are undertaken in a more systematic way while others depend more on the circumstances. The combination makes it difficult to recognize the key issues in survey participation.

The next section focuses on the influence of two essential aspects that often form the basis for the survey participation decision: survey burden and the interaction with interviewer.

2.1.2 Influence of Survey Burden and Interviewers on Non-response

Research in social psychology and survey statistics has shown that non-response could be related to many factors. Factors such as response burden, mode

of data collection, content and design of survey, fieldwork procedures and strategies all have an impact on response rate (de Heer and Moritz, 1997). Some researchers believe that the decision to participate in a survey are likely based on "one or two highly prominent and normally diagnostic considerations" (Groves and Cialdini, 1991), such as the length of the survey or which organization is conducting the survey, while other researchers believe that these factors may interact with each other to affect the participants' decision.

Regardless of the different theories of survey participation, a careful survey design and administration can help to reduce non-response rates. For instance, Richardson, Ampt, and Meyburg (1995) point out that it is very unwise to omit follow-up reminders. Surveys conducted in West Germany and Australia indicate that the use of reminders can significantly increase the number of respondents. Besides survey design and administration, survey burden is another one of the distinctive features that affect survey participation. Currently, there is no standard way to measure survey burden. Survey burden may be defined as the length of the interview. It may also be defined as the number of survey contacts, such as for panel surveys. A common belief in survey research is that survey burden is negatively correlated with survey response (McCarthy and Beckler, 1999; Bogen, 1996). However, previous studies do not consistently support this common belief. Some findings show that shorter interviews yield higher response rate, while some other findings show that longer interviews yield higher response, and some others do not find any obvious pattern (Bogen, 1996). For instance, McCarthy and Beckler (1999) examined various features of survey burden

including length of the survey, time burden of the survey, the number of survey contacts, and the cumulative burden through panel surveys. Their study showed that survey burden on the respondents did not affect future survey cooperation.

The inconsistent findings of the relationship between respondent burden and survey response may be due to the difficulty of isolating the effect of survey burden from the effects of other survey features (Botman and Thornberry, 1992). Surveys are conducted by different organizations and considerable differences may exist among these survey organizations and survey subjects. These differences correlated with respondent burden may be the cause of different response rates. Some studies adopt quantitative analysis approaches other than descriptive statistics and are able to overcome the difficulty and separate the effect of different survey features. Yu and Cooper (1983) analyzed the relationship between interview length (the number of items to be answered) and response rate using linear regression. Their study shows a weak negative correlation between them. A correlation coefficient of r = -0.06 was found. Heberlein and Baumgartner's study (1978) indicated that there is no significant correlation between any of the survey length measures and overall response rate. However, when salience and contacts are controlled, the impact of survey length is significant and longer surveys get lower responses. Their quantitative study showed that one additional question reduced response rate by 0.05%.

Generally the burden of travel surveys depends on the content of the surveys. An on-board transit survey may take a couple of minutes, whereas a face-to-face home interview, such as a travel attitude survey, may take up to an

hour. For household travel surveys, the survey burden depends on how many activities (or trips) a household (or an individual) needs to record. Since trip rates often are the interest of transportation studies, the potential inherent correlation between trip rates and survey non-response cannot be ignored when applying quantitative methods to model survey non-response.

Another important feature of surveys that has an impact on response rate is the characteristics of interviewers. Some researchers believe that the behavior of the interviewer in the household and during interactions with households is critical in the process of obtaining cooperation from potential survey respondents (Groves and Couper, 1998). Nonetheless, relatively few studies have examined the impact of interviewers on survey participation due to the lack of information on interviewers' characteristics. The interviewers' characteristics are not usually collected and thus cannot be studied. Therefore, it is difficult to examine the effects of these characteristics.

The mode of data collection may also have an effect on response rate (de Heer and Moritz, 1997). Travel surveys can be telephone surveys, mail surveys or face-to-face home interviews. Mail surveys are flexible because respondents can choose when to complete the questionnaire. Many household travel surveys are mail surveys that use travel diaries in combination with survey questionnaires. Sometimes phone calls are made for the initial contact and to remind survey participants to return the mail survey. Thus, the difference between mail survey and telephone survey becomes more and more obscure.

Our review of the literature shows that it is generally accepted that survey non-response is affected jointly by a combination of the social environment, the survey design, the interviewers' characteristics, and the attributes of sample units. At the same time, although many studies have been conducted to identify non-response causes, it seems that a clear picture of the causes of non-response has not yet been established. This situation is partially due to the lack of well-defined non-response measures that address different causes of non-response and the lack of information on the decision making process.

There are two main challenges exist in order to identify non-response causes and to achieve a higher response rate. The first challenge is that a large amount of information needs to be collected. In addition to the survey content, extra questions on the decision of participation itself must be included. Other information, such as the interviewer's characteristics and when the interview was conducted, also need to be recorded. Zimowski *et al.* (1997) pointed out that response rates should be reported for each phase of data collection in a multi-phase effort such as household travel surveys. Recording all these data tremendously increases survey cost and survey burden that may have an impact on survey response as well.

The second challenge is the comparability among different surveys. It is necessary to include surveys with different survey features in the study to identify what are the most important features affecting non-response rate. However, people may not behave in the same fashion in responding to different surveys. This may be due to the varying opinions of survey participants on the survey

subject, or simply because of the randomness of the different time points when the surveys are conducted. Part of this problem may be addressed by a controlled experiment, though the cost for the experiments could increase dramatically when more factors need to be monitored. Another solution to the problem is to employ advanced methodology with more interpretational power. In terms of analysis methodology, most research on the topic of survey methods and non-response causes performed descriptive analyses rather than multivariate quantitative studies. The descriptive approach often focuses to compare the response rates at the aggregate level, while the behavior-oriented analysis at the disaggregate level seems able to differentiate various effects on the survey participation decision and provide more significant and consistent insight of non-response.

The next section reviews literature of post-survey bias correction procedures. The review covers the different approaches for item non-response and unit non-response, the quantitative methodology, and some empirical studies.

2.2 POST-SURVEY STATISTICAL APPROACH TO CORRECT FOR NON-RESPONSE BIAS

The advantages of using a post-survey quantitative approach for non-response bias correction are flexibility and avoidance of *ad hoc* methods. When using post-survey statistical approaches to correct for non-response bias, a key concept is the ignorability of the response mechanism. Non-response can be *ignorable* or *non-ignorable*. Non-response is *ignorable* if the non-response pattern does not depend on unobserved characteristics, given the values of observed characteristics. Otherwise, it is *non-ignorable*. For example, if only one

variable y is subject to non-response, the ignorable non-response assumption asserts that non-response is conditionally independent of y given the other observed variables x, shown as follows,

$$P(z \mid x, y) = P(z \mid x), \qquad \text{for } \forall y . \tag{2-1}$$

It should be noticed that the conditional independence is between the non-response behavior and variable y. The ignorable non-response does not require that the non-response behavior be unconditionally independent of y unless y is the only variable.

Without a bias correction process, both ignorable and non-ignorable non-response will lead to inconsistent inference. For ignorable non-response, an appropriate model can prevent misspecifications between endogenous variables and exogenous variables. For non-ignorable non-response, a suitable model procedure can capture the correlations between non-response behavior and endogenous variables due to unobserved characteristics.

In post-survey bias correction procedure, one general model assumption is that survey non-respondents behave the same way as survey respondents. However, this hypothesis is not testable without further validation data. Meanwhile, as many case studies suggest, there are usually systematic differences between respondents and non-respondents and no statistical technique can be relied upon to adjust for all differences. Consequently, many studies point out that it is important to keep non-response rate to a minimum in the first place (Brownstone and Chu, 1997; Horowitz and Manski 1998). Several methods can be used to reduce non-response rate including attempts for a second interview, the

use of advance letters, and proper timing of calls (Richardson, Ampt, and Meyburg, 1995). Nevertheless, with cost and time constraints these methods may not always be successful. Thus, using a statistical compensation procedure is the only remaining approach to correct for non-response bias.

Although the behavioral hypothesis among survey respondents and dropouts cannot be tested without obtaining more information on the dropouts, the upper and lower boundary of estimated parameters can be obtained with certainty. The paper by Horowitz and Manski (1998) discusses how to identify the width of the boundaries on parameters. For example, consider to identify $E[g(y)|x\in A]$, where y is variable of interest, x is an independent variable, and A is the space where x belongs to. Let z =1 indicate that x and y are observed and z = 0 indicate that y is missing, then

$$E[g(y) \mid x \in A] = E[g(y) \mid x \in A, z = 1] * P(z = 1 \mid x \in A) + E[g(y) \mid x \in A, z = 0] * P(z = 0 \mid x \in A). \quad (2\text{-}2)$$

Let $K_0 \equiv inf\, g(y)$ and $K_1 \equiv sup\, g(y)$ for $y\in$ Y, where Y is the domain of y. Then,

$$E[g(y) \mid x \in A, z = 1] * P(z = 1 \mid x \in A) + K_0 * P(z = 0 \mid x \in A) \le E[g(y) \mid x \in A] \le E[g(y) \mid x \in A, z = 1] * P(z = 1 \mid x \in A) + K_1 * P(z = 0 \mid x \in A). \quad (2\text{-}3)$$

Thus, $E[g(y)|x\in A]$ is restricted to an interval of width $(K_1 - K_0) * P(z = 0| x\in A)$.

Because imputation and weighting methods (see Section 2.2.1) are widely used to correct non-response bias, Horowitz and Manski (1998) compared the boundaries of the asymptotic bias of estimates using imputation and weighting methods. The paper concludes that estimates obtained by the weighting method were potentially more biased than those obtained by the imputation method. The

authors found that higher biases in the weighting estimates resulted from a basic flaw in using the weighting method where the weights were not conditional on $x \in A$. Therefore, when the weighting method is employed to correct for non-response bias, it should be applied conditional on the observed attribute $x \in A$. In this case, weighting estimates are not necessarily more biased than imputation estimates.

Another issue that should be noted is initial non-response. Hensher (1993) indicated that initial non-response was selective and should be taken into account. Nevertheless, most of the studies ignore initial non-response bias. In other words, the bias correction procedures are developed conditional on initial non-response. The omission probably occurs because of the difficulty of implementing initial non-response in a model. Generally very little information is known about initial non-response. Thus, it is problematic to model the initial non-response bias. However, this difficulty does not necessarily imply that capturing the initial non-response is not important.

2.2.1 Imputation and Weighting Methods

Imputation and weighting have been described as two major methods of compensating for missing data. Both methods aim to provide reliable survey outcomes and unbiased population estimates. Non-response has two sources: the loss of item information and the loss of unit information. The imputation method is commonly used to deal with missing item information, while the weighting method is commonly used to deal with the loss of unit information.

2.2.1.1 Imputation

The imputation method produces some artificial values to replace the missing data. The method emphasizes the predictive distribution of missing values. It deals with each item or variable individually. There are seven imputation methods to fill in missing values described by Hensher (1987). The first is the deductive method, which simply involves removing the missing item. As the author indicated, this method should not be viewed as a default option, especially when many items are missing, because removing missing items leads to the loss of information and results in a sample coverage problem. The second way is to use a grand sample mean to fill in the missing value. However, it does not make a full use of information obtained from sample. The third method is to use a class mean. In this method samples are divided into different classes based on the values of other variables, such as income, or household size. In each class the class mean is used to fill in the missing values. Although mean imputation is a simple method to implement, it should be noted that the variance of the variable with missing data cannot be consistently estimated by standard variance formulas. The shortcoming is that the mean imputation distorts the distribution of the estimated variable, underestimates the variance and the correlations. The fourth method is the traditional hot-decking method. In this method each non-response variable is replaced by the variable response in the same class that the previous (or sequential) respondent had given. The problem with the traditional hot-decking method is that a single response may be assigned to several non-

responses if a missing value is followed by more than one observation with missing values. The fifth is the modified hot-decking method that minimizes the multiple uses of certain responses by sorting out observations. The sixth imputation method is random imputation that randomly draws a respondent and assigns it to a non-respondent. The last imputation method listed in the paper is to fill in missing values by a regression model. It predicts missing values by estimating an equation between non-response variable and other variables. On the basis of regression imputation, stochastic regression imputation replaces a missing value with a sum of an estimate predicted by regression equation and a residual term which reflects the uncertainty in the predicted variable.

Rubin (1986) proposed a multiple imputation method that imputes the missing item several times. As mentioned before, an important limitation of single imputation method is that standard variance formulas systematically underestimate the variance of estimates. The multiple imputation method replaces each missing value by M ≥ 2 times and the variance can be consistently estimated. When the M sets of imputations are obtained by repeatedly drawing random realizations according to a model that accounts for non-response, the mean of the parameter can be written as

$$\bar{q}_M = \sum_{l=1}^{M} \frac{\hat{q}_l}{M}, \qquad (2\text{-}4)$$

where $\bar{q}_l$ is the estimate based on the l^{th} set of imputations. The combined estimate $\bar{q}_M$ properly reflects the uncertainty in the model that accounts for non-response bias. The variation associated with this estimate has two components: a

within-imputation variance and a between-imputation variance. Combining these two components, the variance of $\bar{q}_M$ can be obtained by

$$\Sigma = \sum_{l=1}^{M} \frac{\tilde{\Sigma}_l}{M} + (1 + \frac{1}{M}) \sum_{l=1}^{M} \frac{(\hat{q}_l - \bar{q}_M)^2}{M - 1}, \qquad (2\text{-}5)$$

where $\tilde{\Sigma}_l$ is the estimator of the variance of $\hat{q}_l$. On the right hand side of equation (2-5), the first term reflects the variance within each imputation and the second term reflects the variance across imputations.

2.2.1.2 Weighting Method

Another approach to deal with non-response bias is by using the weighting method. The weighting method emphasizes predicting the distribution of all responses instead of each missing item. Besides the non-response bias correction, the weighting method can be used for other purposes as well. For instance, if a stratified sampling method is adopted to recruit more rare events in the sample, the weighting method is often used to adjust the survey outcome to provide reliable population estimates. The underlying philosophy of weighting is straightforward. It is primarily used to increase the weight of respondents to compensate for the non-respondents. The basic idea for the weighting method is that a sample unit with a probability of p to be selected is representing p^{-1} units in the population and, consequently, the weight p^{-1} should be used in the estimation.

The general weighting method used in non-response studies can be described as follows. First, a model is estimated to predict the probabilities of survey units responding to a survey. In general, a binary choice model structure is adopted to capture the response process. Then, the inverse of the estimated non-

response probability is used as the weight. One key issue in the weighting adjustment is to obtain a consistent estimate of the response probability. For ignorable non-response, it is assumed that non-response is conditionally independent of the variable of interest. Therefore, the sequential procedure described above would lead to consistent estimates with correct model specifications. However, for non-ignorable non-response, the sequential procedure generally leads to biased estimates. In this case, a simultaneous model system is often adopted for the bias correction (see Heckman's approach in Section 2.2.1.3).

Little and Rubin (1987) pointed out that although the weight method removes non-response bias, it may yield high variance on estimators because respondents with low estimated response rate will receive large non-response weights, which may influence the estimates of means and totals. Meanwhile, weights derived from the inverse of estimated response rate rely heavily on a correct specification of the response model.

Imputation and weighting provide alternative ways in correcting for non-response biases. For panel data, the imputation method can use responses on one wave to predict a missing item response in another wave. Weighting generally does not utilize information across waves. However, cross-wave imputation twists the distribution of changes. Therefore, for analysis focused on individual changes, the weighting method seems to be preferred because the weighting method retains changes exhibited by survey participants in each wave. Next

section reviews some modeling issues which are associated with the application of the weighting method.

2.2.1.3 Modeling Issues Associated with the Weighting Method

Since many behavior variables of interest in transportation studies are discrete rather than continuous, maximum likelihood estimation (MLE) is often used to estimate models instead of the least square estimation. Manski and Lerman (1977) proposed a weighted exogenous sample maximum likelihood estimator (WESMLE) for choice-based samples. The primary objective of their paper is to estimate a discrete choice model when data are collected based on choices rather than random realizations of the population. The WESMLE was initially proposed to obtain a consistent estimation with a choice-based sample. The same methodology may also be applied to panels with sample attrition in order to correct non-response bias when non-response behavior is considered exogenous.

The advantage of the WESMLE is that it can be easily computed to produce a consistent estimate. However, an important assumption of the WESMLE is that the weighting function is exogenously available. With application of modeling non-response behavior, the WESMLE requires that the probability of an individual responding to a survey be explicitly known, or that the response rate can be consistently estimated. The key problem goes back to the response model that predicts the probability of non-response. If non-response

behavior is correlated with the variable of interest, further model specification is needed to obtain consistent estimation.

Heckman (1979) described sample selection bias as a specification error. He incorporated the bias correction procedure within the specification framework. He also presented a tractable computation procedure which allows for the use of simple regression techniques to obtain unbiased estimations. The following two-equation model was used in his paper,

$$Y_{1i} = X_{1i} b_1 + U_{1i}, \tag{2-6}$$

$$Y_{2i}^{*} = X_{2i} b_2 + U_{2i}. \tag{2-7}$$

where Y_{1i} is the variable of interest for individual i, Y_{2i}^{*} is the latent propensity that represents the availability of Y_{1i}, X_{1i} and X_{2i} are vectors of exogenous variables, and β_1 and β_2 are vectors of parameters that need to be estimated. The disturbance terms follow the normal distribution without imposing the IID hypothesis. The assumptions of the disturbance terms are,

$$E(U_{ji}) = 0, \; E(U_{ji} U_{j'i'}) = s_{jj'}, \text{ for } i = i', \; j \neq j' \text{ and } j, j' = 1, 2,$$

$$E(U_{ji} U_{j'i'}) = 0, \text{ for } i \neq i' \; j = j' \text{ and } j, j' = 1, 2. \tag{2-8}$$

Equation (2-6) and equation (2-7) compound the behavior variable of interest with the function that determines the probability of the availability of the variable. The correlation between U_{1i} and U_{2i} indicates that non-response is correlated with the endogenous variable. Based on the estimated parameter, a statistical test can be performed to test the hypothesis that the non-response behavior is non-ignorable.

Suppose that Y_{1i} is observed if $Y_{2i}^{*} > 0$, otherwise Y_{1i} is not observed. Then,

$$E(Y_{1i} \mid X_{1i}, Y_{2i}^{*} > 0) = X_{1i} b_1 + E(U_{1i} \mid U_{2i} > -X_{2i} b_2). \qquad (2\text{-}9)$$

Furthermore, the unconditioned estimates can be derived after transforming the bivariate normal distribution into the probability distribution in one dimensionality. Thus, the bias resulting from non-response is corrected by appropriate model specifications.

The selectivity bias correction procedure originally proposed by Heckman can be extended for studies while the variable of interest is a discrete variable or the analysis is based on panel data. When a discrete variable is of interest, equation (2-6) is transformed to represent the linear relationship among exogenous variables and the propensity of the variable of interest instead of the variable itself. When the method is applied to panel data, more realistic assumptions are imposed on the disturbance terms to account for heterogeneity across time and individuals.

Heckman (1979) also developed a two-step computational procedure to estimate the models. The first step estimates the parameters of b_2 from the binary probit model. In the second step the estimated parameters were transformed and used as a regressor in equation (2-6) to obtain a consistent estimate of b_1. With advances in computing technology and simulation techniques adopted in model estimation, more and more studies adopt a simultaneous estimation procedure.

2.2.2 Empirical Studies

While survey participation theory is extensively studied in social psychology and post-survey methodology is primarily developed in the field of statistics and econometrics, we focus the review of empirical study on non-response in transportation surveys.

Thakuriah *et al.* (1996) applied the Monte Carlo-based simulation method to study non-response effects on logit and gravity models. In their study, for the logit model, non-response was simulated by randomly removing the individuals who had higher transit or higher auto travel costs. For the gravity model, they randomly suppressed some origins or destinations with higher travel time to simulate the non-response process. Their study showed that the parameters estimated from the logit and gravity models were not significantly different. Therefore, the paper suggests that non-response bias has no adverse effect on the parameter estimations as long as the model has been specified correctly. One key factor in this non-response study is the simulated non-response rate. Quantitatively 5% and 30% non-response rates would have different impact on the parameter estimation. The study could be extended to find the critical non-response rate for unbiased parameter estimation.

Kitamura and Bovy (1987) analyzed attrition biases and trip reporting errors for the DMPD. Their paper investigated the relationships among reporting errors in the first wave and second wave, and the attrition process. The methodology proposed by the authors has several advantages. Besides the correlations imposed on the disturbance terms for trip rates and attrition, the

model assumes a chronological dependency across waves. It is assumed that the correlation coefficient was a function of exogenous variables. It is found in the study that households with older children and more vehicles tend to decline the survey request for the second wave. In addition, less-educated households are less likely to participate, too. Besides household demographic variables, this study probably is the first to test the relationship among trip frequency and panel attrition in a rigorous statistical mechanism. They found that households with a larger expected number of trips per person in the first wave are more likely to continue in the second wave.

Another study using data also collected in Netherlands seems to suggest that higher attrition rates are related to higher mobility (Arentze ***et al.***, 2000). The model results indicate that the higher the number of trips in the first wave, the more likely people will drop out. However, the estimated coefficient for the number of trips is not significant with a *t*-value of 0.78.

Chung and Goulias (1995) examined two sources of sample selection bias, panel attrition and residential relocation in first two waves of PSTP, on the frequency of activity participation in the second wave. The methodology follows Heckman's approach and the model was estimated using Heckman's two-step computational procedure. In terms of attrition behavior, they found that households with a higher car ownership, more workers, and longer duration of residence in wave one are more likely to participate in both waves. In addition, low-income, single-adult, childless, and younger households tend to stop responding after the first wave. These results are consistent with previous studies.

The results for residential relocation are a bit interesting. It is found that residential tenure has a negative impact on the survey participation for the second wave. Households living in the current residence for five or more years are less likely to continue participating in the survey for the second wave.

The unexpected results of residential relocation might be a combinatorial result of data collection problem and the basic behavioral assumption on non-respondents. The key issue seems to be whether or not households are traceable once they change their residence location. If not all moved households are reached by a survey request for the second wave and automatically these unreachable households are considered as non-respondents, the basic assumption that respondents and non-respondents would behave in the same way in terms of residential relocation then can not be held any more. It seems that more data collection and information recording efforts need to be carried out to address this limitation.

Sen *et al.* (1995) estimated a logit model to identify response rates for key population subgroups based on a household travel survey for Chicago area transportation study. Their findings support a common hypothesis that the lower-income, less-educated households are systematically underrepresented because they are more likely to decline the survey request. On the other hand, there are some different findings compared to Chung and Goulias' study. For instance, the model suggests that larger households (with four or more members) are less likely to respond to the survey and households with no vehicle have higher response rate. Opposite findings are revealed in the PSTP. The Chicago subway system

may account for the higher response rates among households with no vehicle. The similarities and dissimilarities in the results of these two studies suggest that common points can be observed in participation behavior toward different travel surveys, at the same time the behavior varies under different social environment.

Pendyala and Kitamura (1997) proposed a weighting method for attrition in choice-based panels and applied the method to the PSTP. The weighting method they proposed corrected for two types of biases. The first bias was due to the attrition. The second bias resulted from non-randomly selected choice-based sampling itself. The authors also developed a weighting scheme for random refreshment samples for the panel. To accommodate the bias of a choice-based sample, they followed the weighting method proposed by Lancaster and Imbens (1990):

$$w(j) = \left[\sum_{j \in C_n, n \in N} \frac{H(n)}{Q(n \mid l)} \right]^{-1} \qquad (2\text{-}10)$$

where $H(n)$ represents sample probability of the n^{th} choice stratum, $Q(n \mid l)$ is the population proportion of the n^{th} choice stratum given parameter λ that describes choice phenomenon, and C_n stands for the choice set of the n^{th} choice stratum. To compute the joint weight, the study estimates a bivariate probit model for initial choice and attrition behavior and found that, in the PSTP, the initial choice behavior is independent of attrition behavior. This result indicates that the sequential modeling procedure is able to generate consistent estimates.

Brownstone and Chu (1997) combined the multiple imputation method and the WESMLE to study the dynamic mode choice pattern by using the Southern California Transportation Panel Data. The paper first estimated a

binomial logit model to capture the attrition process. The attrition probability was assumed to be independent of the mode choice. The estimated model parameters were assumed to be normally distributed. Second, the random realizations of the parameters were drawn based on the estimated mean and standard deviation values of the parameters. Consequently, the realizations of the attrition probability were computed. Third, using the inverse of the attrition probabilities as weights in the WESMLE, a multinomial logit mode choice model was estimated for mode choice. This procedure was repeated for m times and m attrition probabilities were obtained for each record. The parameters in the multinomial logit model were estimated m times. Finally, following equation (2-3) and equation (2-4) proposed by Rubin (1986) for the multiple imputation method, the mean and the standard deviation of the parameters of the multinomial logit model were computed. The study reveals similar insight in attrition behavior as other studies except that the estimated parameter suggests people with more than three vehicles are less likely to attrite.

2.3 ATTRITION IN MULTI-WAVE PANEL DATA

The empirical studies reviewed in the previous section mainly focused on two waves of panel data. When considering attrition in multi-wave panels, two specification issues need to be addressed besides common modeling considerations for panel data. One is the impact of lagged variables. The other is the effect of the current spell of participation. In some models based on the *missing at random* assumption (MAR, Rubin, 1976; Little and Rubin, 1987), the

probability of attrition depends on lagged but not contemporaneous variables, while the model proposed by Hauseman and Wise (1979) allows the probability of attrition to depend on contemporaneous but not lagged variables. A more realistic hypothesis test is that panel attrition is a joint result of both lagged and contemporaneous variables and it should be reflected in the model specification. The other modeling issue is to properly incorporate the impact of participation duration. Duration dependence can not be overlooked when modeling panel attrition because survey units are repeatedly subject to non-response and their past experiences during the survey often play an important role in the decision of repeatedly participation.

Probably due to the complexity in model specification and computational burden, little literature has focused on dealing with multi-wave panel attrition and quantifying its effect on the variable of interest, especially in transportation studies. We found a few studies done by econometricians and psychology researchers. Ridder (1990) proposed a general model structure for attrition procedure in multi-wave panel data. He described a model system that included models for the variables of interest and models for attrition probability for all waves. All models are conventional random effects models. The model structure for the variable of interest is written as,

$$y_{it} = b' X_{it} + a_i + e_{it}, \quad i = 1,...,N \text{ and } t = 1,...,T, \qquad (2\text{-}11)$$

where y_{it} is the variable of interest, X_{it} is a vector of exogenous variables, α_i and ε_{it} are the error terms, and β is a vector of parameters to estimate. The models for attrition behavior is written as,

$$a^{*}_{it} = g_0 'w_{it} + g_1 y_{it} + g_2 a_{i,t-1} + ... + g_t a_{i,1} + g_{t+1} d_{it} + g_{t+2}(1 - a_{i,t-1}) + d_i + h_{it}$$

$$a_{it} = 1 \text{ if } a_{it}^{*} \geq 0$$

$$a_{it} = 0 \text{ if } a_{it}^{*} \leq 0$$

$$i = 1,...,N \text{ and } t = 1,...,T, \qquad \text{(2-12)}$$

where $d_{it} = \sum_{k=1}^{t-1} \prod_{s=1}^{k} a_{i,t-s}$.

Ridder considered various factors that have potential impacts on attrition in the model. The latent propensity of attrition in period *t*, a_{it}^{*}, is a function of exogenous variables w_{it}, the endogenous variable y_i in period *t*, the individual's attendance in all previous periods a_{it}, and the length of the current spell of participation d_{it}. In equation (2-12), all the γ's are the parameters to be estimated, d_i and h_{it} are the error terms. The model has some advantages in capturing the characteristics of attrition behavior in panels. For example, the model can be used to estimate initial attrition, and it allows for the return of individuals who have left the panel at an earlier date. However, the paper does not apply the model to an empirical analysis mainly due to the difficulties in estimation. The computational difficulty of this model remains unknown.

Taris (1996) studied non-response in multi-wave panel data using discrete-time Markov chain models. The model considers non-response and dropout as a result of a Markov process, assuming that the waves are equally spaced in time. The study focuses more on the attrition pattern across waves rather than individual characteristics. A small empirical example is demonstrated in the paper using three-wave panel data from the social integration of young adults. The rejection of the first order Markov chain model (with a constant response

rate) implies that the attrition is not constant over waves. The paper proposed a mixed model that consists of two chains for "stayer" (people who always respond to the surveys) and "mover" (people who miss some surveys) respectively. Not surprisingly, the mixed model produces a better fit to the data.

Nijman and Verbeek (1992) conducted various statistical tests to examine the impact of non-response on estimates of a life cycle consumption function using monthly collected Expenditure Index Panel of Netherlands during the period of April 1984 to March 1987. Their tests suggest that there might be an attrition problem, but the evidence is not decisive. Besides, some tests heavily depend on the assumption of response behavior mechanism. The paper recommends the use of simple procedure to test attrition bias before adopting a computationally demanding general model structure.

Ridder (1992) applied the Hauseman-Wise model to analyze four waves of DMPD. He found that the model does not perform as expected probably due to the incorrect distributional assumptions. The restrictions on the correlations between individual effects and disturbances may even prevent the detection of the non-random attrition. He also found that non-random attrition does affect the time varying coefficients of the trip frequency model but not the time constant regression coefficients.

2.4 SUMMARY

The literature review indicates that it is essential to reduce non-response through effective survey design. Both social psychologists and statisticians have

been trying to understand survey participation behavior from different view points. It is generally believed that non-response is a combinatorial result of social environment, survey features, sample attributes, and interviewer's characteristics while some researchers pointed out that survey participation decisions are often made on the basis of one or two highly prominent factors.

Common approaches adopted in earlier transportation studies to accommodate attrition bias are the weighting adjustment and a model-based method. A binary discrete choice model is widely used to model non-response/attrition process. In panel studies, two-wave data are used in most of the studies. Few studies correct for attrition bias in multi-wave panel data. It seems that there is no clear-cut answer to the question of how to handle non-response in panel data. Some theoretical model structures have been extensively discussed among econometricians. However, the computationally intensive estimation procedure precludes its application in empirical analysis.

The next chapter will discuss the research approach adopted for this dissertation based on the review of existing literature.

Chapter 3 Research Approach

In the past two decades there has been a significant increase in the attention paid to non-response in the survey literature. There are continuous calls for advanced quantitative analyses in order to build and test theories of survey participation, to evaluate causes of survey participation suitable for both respondent and non-respondent cases, to construct similar measurement for survey-specific cases, and to estimate bias adjustment models. This dissertation attempts to provide a comprehensive analysis on the subject. The work presented in this dissertation is developed in two stages. In the first stage the variable of interest is panel attrition alone. The goal is to build a model compatible with survey participation behavior that is computationally tractable as well. The second stage focuses on the potential correlation between panel attrition and trip frequency, which is always of interest to travel behavior analysts. This chapter discusses the conceptual and theoretic framework for the analysis of panel attrition and the contents of survey. Section 3.1 presents a conventional approach, i.e., discrete choice model, used to model multi-wave panel attritions and summarizes the practical barriers that prevent wider applications of this approach. These application limitations lead us to adopt a hazard-based duration model structure for panel attrition modeling. Section 3.2 discusses the potential impact of travel behavior on non-response in household travel surveys.

3.1 MODELING DISCRETE VARIABLES IN PANEL DATA

In earlier literature, the sample unit's participation decision is often modeled as a discrete variable. It is viewed as a choice among available alternatives. In the case of survey participation, the available alternatives are either "refuse" or "respond" to a survey request. The general approach to analyze choice behavior is discrete choice model. The standard model structure has been applied to a variety of disciplines because of its simplistic behavioral assumptions. Meanwhile, it is recognized that there are gaps between the complex behavioral theory and its simplistic representation in the model formulation, and tremendous efforts have been undertaken to close the gap.

3.1.1 Discrete Choice Model and Choice Behavioral Theory

Discrete choice model has its foundation in the theory of random utility maximization. The concept of random utility theory was first introduced by mathematical psychologists (Marschak, 1960; Luce, 1959) and later was transformed into a form suitable for econometric applications by economists (McFadden, 1968, 1975, 2000). The original formulation of random utility maximization follows economic consumer theory. The basic assumptions for discrete choice models are:

1. The available alternatives are finite and mutually exclusive;
2. Each alternative i is associated with a net utility U_{iq} for individual q,
3. Individuals possess perfect information, act rationally as decision makers and always select the alternative that has the maximum net utility.

The net utility U_{iq} consists of two components: a measurable, systematic component V_{iq} and a random disturbance term e_{iq}. Thus,

$$U_{iq} = V_{iq} + e_{iq}, \tag{3-1}$$

where V_{iq} is a function of observed individual and alternative attributes and e_{iq} is a disturbance term representing measurement and observation errors, as well as unobserved factors such as habit and taste variation in preference among individuals. For computational reasons, the disturbance term is assumed to follow different probability distributions that are associated with various forms of the model (e.g., normal distributions for probit models and Gumbel distributions for logit models).

The simplified assumptions lead to a tractable structure of discrete choice models and its broad application in travel demand analysis since the 1970s. Nonetheless, there are still gaps between discrete choice models and choice behavioral theory. For example, the property of Independency from Irrelevant Alternatives (IIA) in multinomial logit models is not always compatible with reality. The gap may be due in part to the contradictory nature of mathematical formulation and human behavior. Mathematical models are rigorous and often in a one-to-one input-output format, while human behavior is often driven by beliefs and attitudes, and consequently the motives can hardly be detected in many occasions. Furthermore, the strong assumptions imposed on discrete choice models intensify this intrinsic difference and lead to the inconsistency between discrete choice models and behavior theory.

Figure 3-1 shows a decision-making process, where the darker arrows describe the process corresponding to the economic point of view and the lighter arrows link the psychological elements involved in decision-making (McFadden, 2000). The diagram indicates that the concepts of perception, preference, and attitude appear in both economic and psychological views but work in different ways. The economic view focuses on transforming information into measurable attributes and making a decision based on some rules, while the psychological view seems to emphasize on the interactions among these elements.

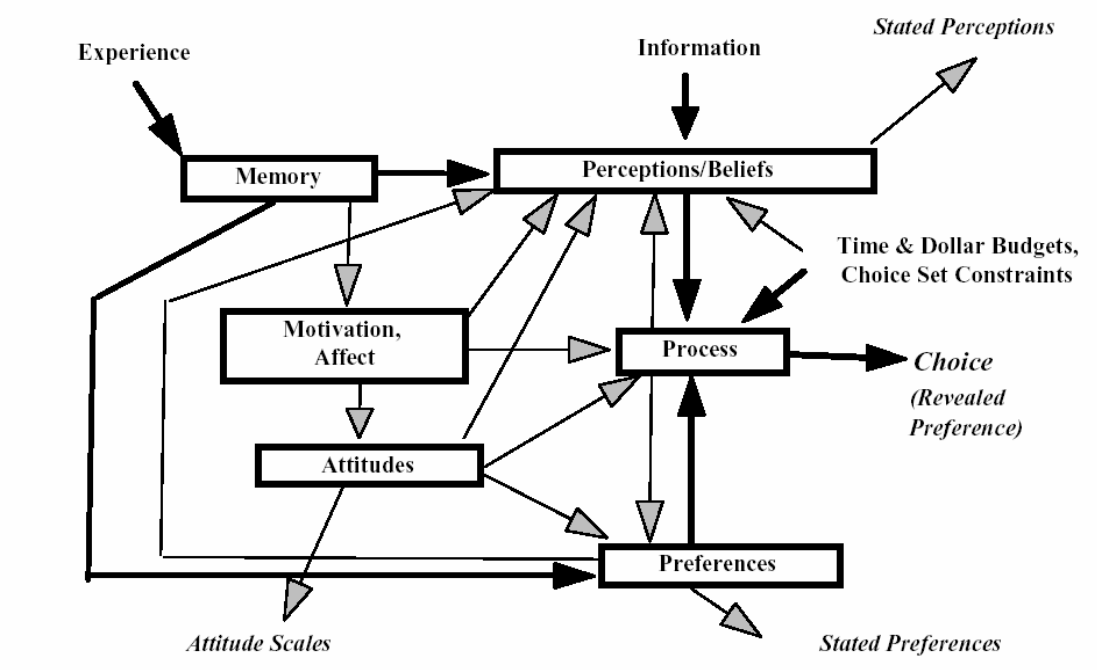

Figure 3-1: Elements in decision-making process (cited from McFadden,2000)

The understanding of the decision-making process undoubtedly helps to narrow the gap between discrete choice models and the reality of choice behavior. The contents of discrete choice models have been enriched dramatically since the early 1990s. One main limitation of the standard discrete choice model structure is the implementation of unobserved psychological elements. The extended model structures all target a more realistic interpretation of choice behavior. Most of them enhance the specification of the disturbance term in one way or another to account for unobserved heterogeneity and preference variation across individuals. A distinctive example of a recent development is mixed logit model, which embodies a flexible structure (McFadden, 2000; Bhat, 2001; Train, 2002). The disturbance term in mixed logit model is at two levels. Level one includes a disturbance term following a Gumbel distribution. The standard distribution assumption reduces the dimensionality of the integral involved in choice probability computation and eases the estimation. Level two imposes no restrictions on the disturbance term, which provides the flexibility. Another structural extension is the integrated choice and latent variable models that characterize psychological elements as an observed indicator and incorporate it into the choice model (Ben-Akiva, ***et al***., 2001).

As one of the objectives of this dissertation is to understand the survey participation behavior, a flexible disturbance structure seems essential with imperfect information. In addition, survey burden is quantified as an indicator that is considered as a latent variable and incorporated in survey participation choice.

As a flexible disturbance structure presents a realistic interpretation in choice behavior in cross-sectional data, the behavioral mechanism is more complex when a decision-maker faces repeated choices in panel data. A general formulation to model discrete variables in panel data is presented in the next section. The application limitations of this general structure result in an innovative approach to model survey participation decisions in panel surveys.

3.1.2 Choice Behavior in Panel Data

Panel data are advocated in behavior studies to analyze the impact of past experience and newly perceived information. In practice, many panels are collected in more than two waves. Participation in panel surveys is a learning process for sample units. The learning factor may be a constant or may vary over time. It is even more difficult to guess how this learning process would affect the repeated participation decision that a household faces. Because survey participants have experienced the survey and their experience accumulates, their attrition behavior after two waves could be substantially different from that after one wave. In addition, because of the learning process and the presence of state dependency, it would be more appropriate to model panel attrition across waves simultaneously instead of wave by wave separately.

When modeling attrition patterns in multi-wave panels, the key issue is to capture the correlations among attritions in the sequential waves. Because survey participants repeatedly attend surveys that share the same subjects, use similar data collection methods, and require comparable amount of time to complete, the

past experience of the survey surely affects the decision to continue responding to the survey. The impact of the past experience can be reflected in the attrition model in different ways. The model may show that a current decision explicitly depends on a previous decision, or the correlation may exist due to unobserved factors. Previous studies (Kitamura and Bovy, 1987; Arentze *et .al*, 2000) have used binary discrete choice models to model the attrition patterns after the first wave. A similar methodology could be repetitively applied to model attritions in the following waves without considering the effect of the past survey experience. However, this approach ignores the sequential character of panel surveys and the ignorance may lead to inconsistent estimations and biased conclusions.

This dissertation focuses on modeling survey participation decision in multi-wave panel data. Heckman (1981) provided a general form for discrete panel data analysis. In his general model, no restrictions are imposed on the disturbance term. The only necessary condition is that the variance of the disturbance term be a positive definite covariance matrix. The relaxation of the IID assumption on the disturbance term allows models to capture unobserved heterogeneity across individuals and time. In addition, four types of effects on the utility function are presented in the general form. The first one is the effect of exogenous variables on the current utility. The exogenous variables may include past exogenous variables, current exogenous variables, and expectations of future exogenous variables that might determine current choice. The second effect is the effect of the entire past history of the process on the current choice. The third is the cumulative effect of the most recent continuous experience in a state on

current choice, which can be interpreted as duration dependence. The second and third effects provide the key features of state dependence. The fourth is the effect of habit persistence which is implemented in the form so that prior utility to select a state instead of prior occupancy of a state determines the current choice probability.

Although the general form proposed by Heckman (1981) is sufficiently flexible, some application issues deserve consideration. One issue is the numerous specification tests. Consider the first effect implemented in the general form. The exogenous variables could be past, current, and future variables for multiple waves. The number of variables may be in hundreds and thousands. Modelers' judgment would be a dominant source of model specification.

Another application challenge is to estimate the parameters without running into computational difficulties. For instance, consider a two-dimensional correlation imposed on the distribution of the disturbance terms for a panel data study. The correlation can occur across time when decisions are repeatedly made by the same individual. Similarly, the correlation may exist across individuals who make decisions at the same time. In fact distributions with higher dimensionality can be imposed on disturbance terms. The computational difficulty of estimating these models with a relaxed disturbance term structure lies in the multi-dimensional integral that needs to be calculated to compute choice probabilities. One method to avoid multi-dimensional integrals and relieve the computational burden is to generate auxiliary variables and adopt two-stage estimation procedures. The limitation of the two-stage estimation is that it

heavily relies on the mathematical property of distribution and correlation that is imposed on the disturbance term. It is unlikely that a two-stage estimation procedure can fulfill the needs in a multi-wave panel study. The reduction in dimensionality of the disturbance term distribution often requires a multi-stage procedure and it is difficult to derive auxiliary variables in the multiple stage estimation procedure. In addition, the estimates obtained from two-stage procedures are not efficient.

To obtain consistent and efficient estimates, discrete models are usually estimated using full information maximum likelihood estimator. In the full information maximum likelihood estimation, one approach to accommodate the multi-dimensional integral is to use Monte Carlo simulation. In the simulation, a number of realizations of disturbance terms are randomly drawn and used to compute choice probabilities. The average log-likelihood function is then computed based on these random realizations. The parameters are obtained in maximizing the average log-likelihood function. The estimated parameters are consistent, asymptotically efficient, and asymptotically normal under weak conditions. The asymptotic property of the estimator requires a large number of random draws in the simulation. However, the large number of random realizations usually makes convergence slow. Another approach to ease the computation difficulty is to draw the random realizations in an "intelligent" fashion. Recently the quasi-Monte Carlo simulation techniques have been applied to evaluate multi-dimensional integrals involved in the log-likelihood function (Bhat, 1999; Train, 1999). Bhat (1999) compared Monte Carlo and quasi-Monte

Carlo simulation methods in estimating the mixed logit model. The results showed a substantial reduction in computational cost with superior accuracy when using quasi-Monte Carlo simulation. Train (1999) confirmed the considerable reduction in computational time. Therefore, the quasi-Monte Carlo method is a much more powerful tool for the estimation of complex choice models. The quasi-Monte Carlo method is used to estimate the models proposed in this dissertation.

As much in the same way as advanced estimation techniques ease the computational difficulties, computational cost can also be reduced by choosing the appropriate model structure. For example, one can consider a state dependence study. The state dependence exists when the current decision is affected by the past decision history. In this case the current decision and the past decision history are all endogenous variables that need to be estimated. Simultaneous estimation could be tedious. However, if the modeling subject is the conditional probability given the past decision history instead of the unconditional probabilities of sequential decisions, the model structure can be simplified. Another advantage that has not been fully taken in modeling survey participation is the binary feature of the decision itself. Furthermore, it is common practice that once a household drops out, it is not likely to return in the future waves. Then, survey participation can be considered as a continuous process with discrete end points. These characteristics of survey response indicate that an important period in the past history that matters to the decision is

the waves a household has participated in. The state dependence across waves is mainly the duration dependence.

When the number of waves in which a household would participate is the subject of interest, a hazard-based duration model seems more suitable than a discrete choice model. The conventional discrete choice approach views the participation decision at a point in time repeatedly for panel surveys, while the hazard-based duration model views this process along the time scale from the initial wave across waves. The model structure has the advantage of implementing duration dependence without too much computational trouble in model estimation. Meanwhile, time-varying covariates can be conveniently incorporated for specification tests. Therefore, the hazard-based duration model structure is adopted in this dissertation to model non-response behavior in multi-wave panel data.

3.2 TRAVEL BEHAVIOR AND NON-RESPONSE IN HOUSEHOLD TRAVEL SURVEY

As pointed out in the previous chapter, non-response behavior is often associated with the survey subject in one way or another. Household travel surveys collect information on when, where, why, and how individuals make trips in a conventional paper-pen format. The way a household travels may affect their decision to participate in the survey. There seems lack of hard proof about which facet of travel activity has more influence on survey response. Some transportation literature has investigated the relationship between non-response

and travel mode and found no correlation between them (Pendyala and Kitamura, 1997).

Theories on survey participation behavior suggest that survey burden is a major consideration in the decision-making process. In household travel surveys eligible household members are asked to record every travel activity in a travel diary. Generally, the individuals need to report travel time, origin and destination places, travel mode, and the number of travel partners if there are any. The workload to report each trip is about the same. Thus, survey burden can be well represented by trip frequency and the number of trips made by a household during the survey period can potentially determine the participation decision.

Trip frequency is also an important concept in the context of travel behavior analysis. The question of how many trips are made is the first one addressed in the conventional urban transportation modeling system. Travel activity, along with other daily activities, is also of interest in the activity modeling approach. A consistent estimate of how many trips are made is the core of transportation modeling. Therefore, it is necessary to consider survey participation and trip making simultaneously. The study is valuable to portray a full picture of non-response and travel behavior.

When survey participation and trip frequency are both of interest in panel data, much of the modeling effort focuses on how to bond one with the other, besides the issues discussed in Section 3.1. The potential correlation of trip frequency and survey participation may be exhibited in a direct or indirect way. Reflected in model structure, the direct effect can be expressed in a form of $y =$

$f(x)$. The indirect effect is often implemented by introducing a correlation in disturbance terms. Considering these direct and indirect impacts, three hypotheses will be tested in the model system:

- More trips made by a household during the survey period will directly result in the household quitting the survey;
- The variation in trip frequency and survey participation decision can be accommodated by the observed households demographic variables, such as household size and household income;
- The correlation between trip frequency and survey participation can only be captured by a common unobserved factor.

Detailed issues about model specification to test these hypotheses are discussed in Chapter 6.

3.3 SUMMARY

This chapter discusses the research approach adopted in this dissertation. The models are developed for two purposes. The first goal aims for a better understanding of panel attrition. In addition to factors in the social environment, survey design, and individual characteristics, intuition suggests that duration dependence cannot be overlooked when modeling the survey participation decision in multi-wave panel data. The hazard-based duration model structure is convenient to accommodate duration dependence and is selected to model the survey participation duration in panel data. The second model is to detect the potential correlation existing in survey response and travel behavior. This model

examines various impacts on survey participation, and also provides consistent and efficient estimates of trip frequency.

The next chapter introduces the data, seven waves of the PSTP, which are used for the empirical analysis, followed by a brief description of the data collection procedure, and descriptive statistics of household demographics and travel features.

Chapter 4 Data Description

The models proposed in this dissertation are estimated using Puget Sound Transportation Panel (PSTP). This chapter provides a brief introduction on data collection procedure, as well as the descriptive statistics for sample evolution, household attributes, and trip characteristics.

4.1 INTRODUCTION

The PSTP was collected for three purposes: to monitor changes in household demographics, to monitor changes in travel behavior and responses to changes in the transportation environment, and to investigate the effects of changes in attitudes and values on travel behavior (PSRC, 1997). To monitor changes over a period of time, the panel survey recruited panel participants in each of three household subgroups. These household subgroups are:

- Households without regular transit users or carpoolers
- Households with regular transit users
- Households with regular carpool commuters.

The regular transit users are those who have at least four one-way trips per week and the regular carpool commuters are those who share ride for work trips.

The data were collected using two-phase surveys. The preliminary phase was carried out by telephone. The sample households were selected through a random-digit dialing process. However, the initial household samples did not produce an adequate sub-sample of transit users and carpoolers for the analysis.

To increase the shares of transit users and carpoolers in the survey, extra household samples were obtained by re-contacting respondents from other transit surveys who had agreed to participate in a future study and by distributing letters on randomly selected bus routes for volunteers. After the preliminary phase determined the sample households' eligibility and confirmed their cooperation in survey participation, the main phase was carried out through mail. A cover letter, a household questionnaire and travel diaries were sent out to the selected households. The households were solicited to provide information on household demographics as well as a two-day travel diary for each survey period.

The first wave of the panel was initiated in 1989 with 1712 households returning completed diaries. The following six waves were collected from 1990 to 1997 at approximately one year intervals. As attrition did take place, the following waves were refreshed with new household samples. The refreshment samples were drawn for two purposes: to keep the sample size at appropriate level and to maintain representativeness of the population. As Hensher (1987) pointed out, the representativeness can be interpreted as representing the population from which the original data were drawn. But it can also refer to reflecting changes in population characteristics over time. Since one of the objectives of the PSTP is to monitor changes in household demographics, the dynamics in the population should be presented in the sample.

As a result, the approach used in the replacement sampling was to replicate the original panel as close as possible. In addition, newly migrated households were recruited to represent the changes in the population. The data of

refreshment households were also collected through two-stage surveys. First, households were accessed, screened, and qualified through the random digit dialing. One of the criteria for the refreshment qualification is to match geographic locations and travel modes with the survey dropouts. The household demographics, such as the household life-cycle stage, were also consistently monitored. Then, the travel diaries and household demographic data were collected through mail. For the sampling of newly migrated households, initially the households were drawn from new residential customer lists from a major telephone company and a major electric utility company, as well as from the random digit dialing. However, the lists provided by these companies were far from complete. It turned out that most of newly migrated households were drawn from the random digit dialing.

For each wave, three sets of data were provided by the PSRC. These data are household demographic, person demographic, and travel activity data. The household demographic data contains information such as household size and household income. The demographic information of each household member who filled out the travel diary (age 15 or older) was recorded in the person data. The travel data consists of travel activities conducted by each eligible household member during the two-day survey period. The activity data provides information such as the origin and destination of each trip, travel mode, departure time, and other trip-related characteristics.

The rest of this chapter is organized as follows. Section 4.2 describes the sample evolution of the PSTP. Section 4.3 explains the imputation methods used

to impute demographic and trip characteristics variables for the households who provided partial information to the survey. Section 4.4 provides the descriptive statistics on household demographics, followed by a section illustrating the changes in trip frequency and travel mode across the seven waves.

4.2 SAMPLE EVOLUTION

In this study, households who returned completed or partially completed survey questionnaires and travel diaries are referred to as survey participants. Unlike some other studies that did not consider those who provided incomplete information as survey participants, we simply attempt to maintain every piece of information and to avoid further reducing the sample size. Another reason for including partially completed sample units in the analysis is that one of the primary objectives of this study is to analyze non-response behavior of households in a transportation panel survey, unfinished survey questionnaires may be a good indicator of non-response in future surveys. Loosveldt, Pickery, and Billiet's study (1999) has shown that item non-response variables are significant predictors of unit non-response for the Belgian General Election Study. Therefore, in our study, the households who returned incomplete diaries are not removed from the data set. These households are referred to as survey participants with missing data in the remaining contents of the dissertation. Different imputation methods are applied to these households to impute the missing items in the pre-analysis stage. The imputation methods will be discussed in the next section.

Figure 4-1 illustrates the sample composition for each wave. As the samples were stratified on the basis of travel mode, about 67% households were drawn from the single-occupancy-vehicle (SOV) user group for the initial wave. The households with regular transit users and carpoolers compose 22% and 11% of the sample respectively for wave one. The sample stratifications were managed not deviating much from these percentage levels for the following waves. Figure 4-1 also shows the sample composition in terms of when the households first entered the survey. In the first wave, there were a total of 1712 households who returned travel diaries. In wave two, there were a total of 400 new recruited households which is about 20% of entire survey participants. So were the following wave three and wave five, in which the new recruited households consist of roughly 20% of the entire sample. However, in wave four, wave six and wave seven, the newly recruited households are a much higher percentage (27%, 42% and 33% respectively) of the survey participants. With sample size carefully managed at a comparable level, more newly recruited households means fewer continuous survey participants. One possible reason for higher dropout rates may be that an attitude survey was conducted with the travel diary survey for these waves. The attitude survey may bring extra survey burden and therefore cause more households not to return the survey forms. To test this hypothesis, a dummy variable representing an attitude survey conducted in the same year is later included in the panel attrition model.

Table 4-1 demonstrates the frequency of dropout households. The households are segmented by the wave when they initially entered the survey.

Figure 4-2 shows the corresponding percentage figures. The diagram in Figure 4-2 indicates a clear pattern of households' nonresponse behavior in time sequence. Many households stopped responding to the survey after staying in the survey for one or two waves. The dropout rate reaches its highest point after households had been with the survey for two waves except for the households who entered the survey in wave 4. Afterward, the dropout rates declined as the households had been responding to the survey for three or more waves. The descriptive statistics indicates that two waves might be considered as the fatigue duration for the participants in this case.

Figure 4-3 reveals the household survey participation durations. In this plot, the shadowed columns represent the number of households who were censored (i.e., the households responded to the last wave of the survey, wave 7), while the blank columns represent the number of households who were not censored. There are no clear patterns that can be observed from this diagram. In general, the majority of the households stayed with the survey for either one or two waves. Not including the censored cases, 34.7% of the total households attended one wave of the survey and 26.9% attended two waves. The number of households who continued staying in the survey for more than two waves is substantially reduced. A total of 488 households, which is 9.6% of the total sample, completed the survey for three waves. Even fewer households responded to more than three waves.

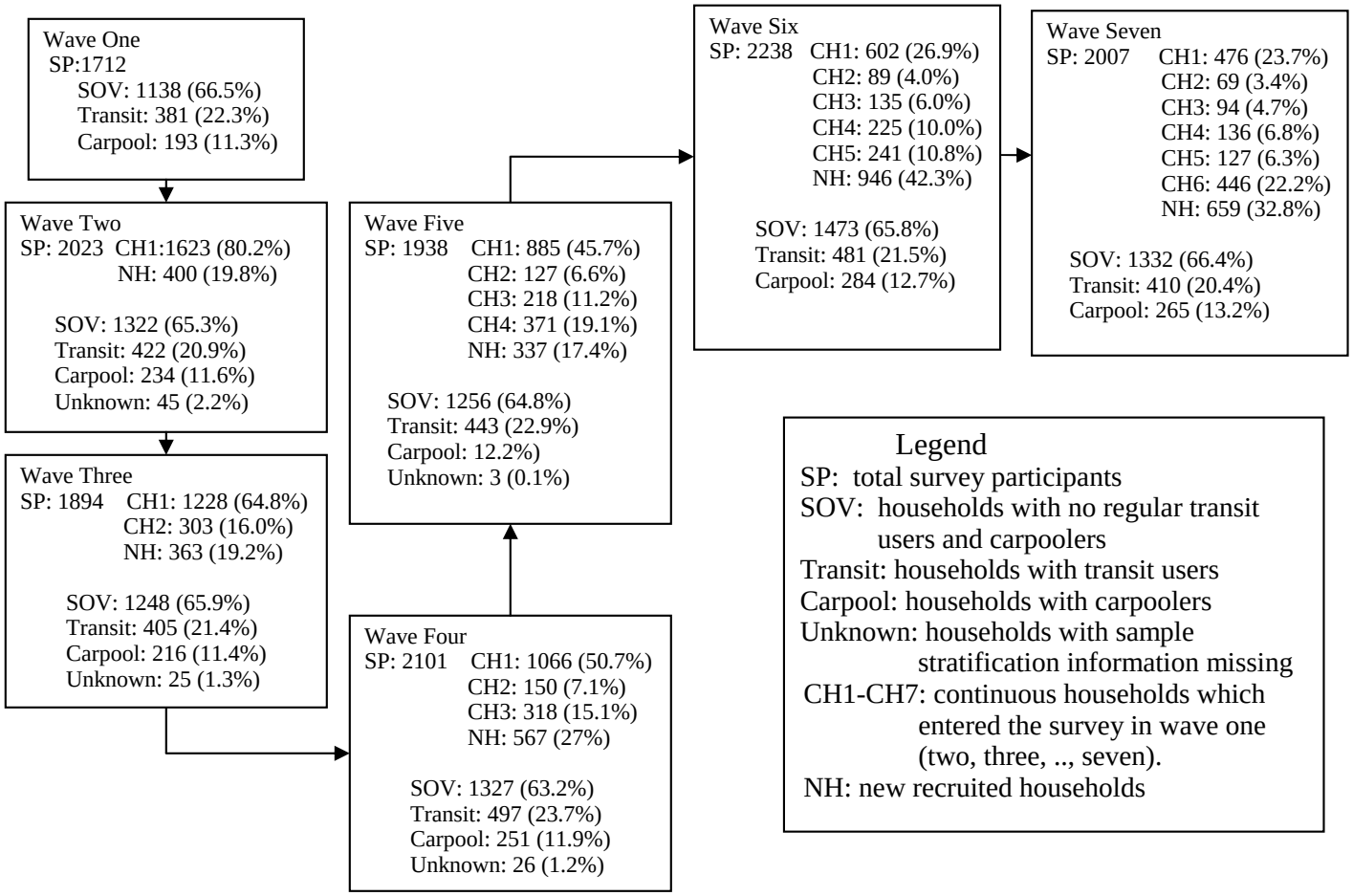

Figure 4-1: Sample stratifications of the PSTP

Table 4-1: Frequency of dropout households

Starting wave	New households	Dropouts after one wave	Dropouts after two waves	Dropouts after three waves	Dropouts after four waves	Dropouts after five waves	Dropouts after six waves	Households in wave seven
Wave one	1712	89 5.2%	395 23.1%	162 9.5%	181 10.6%	283 16.5%	126 7.4%	476 27.8%
Wave two	400	97 24.2%	153 38.2%	23 5.8%	38 9.5%	20 5%	-	69 17.2%
Wave three	363	45 12.4%	100 27.5%	83 22.9%	41 11.3%	-	-	94 25.9%
Wave four	567	196 34.6%	146 25.7%	89 15.7%	-	-	-	136 24.0%
Wave five	337	96 28.5%	114 33.8%	-	-	-	-	127 37.7%
Wave six	946	500 52.9%	-	-	-	-	-	446 47.1%
Wave seven	659	-	-	-	-	-	-	659 100%
Split households	110	86 77.7%	16 15.2%	4 3.6%	2 1.8%	1 0.9%	1 0.9%	0

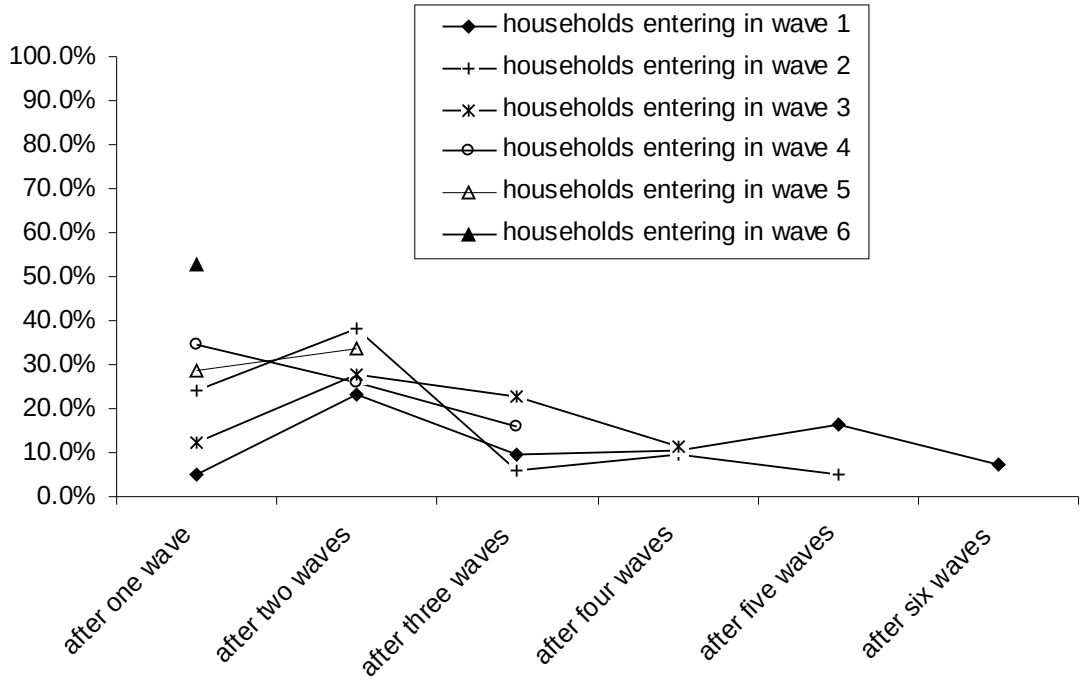

Figure 4-2: Household dropout percentage

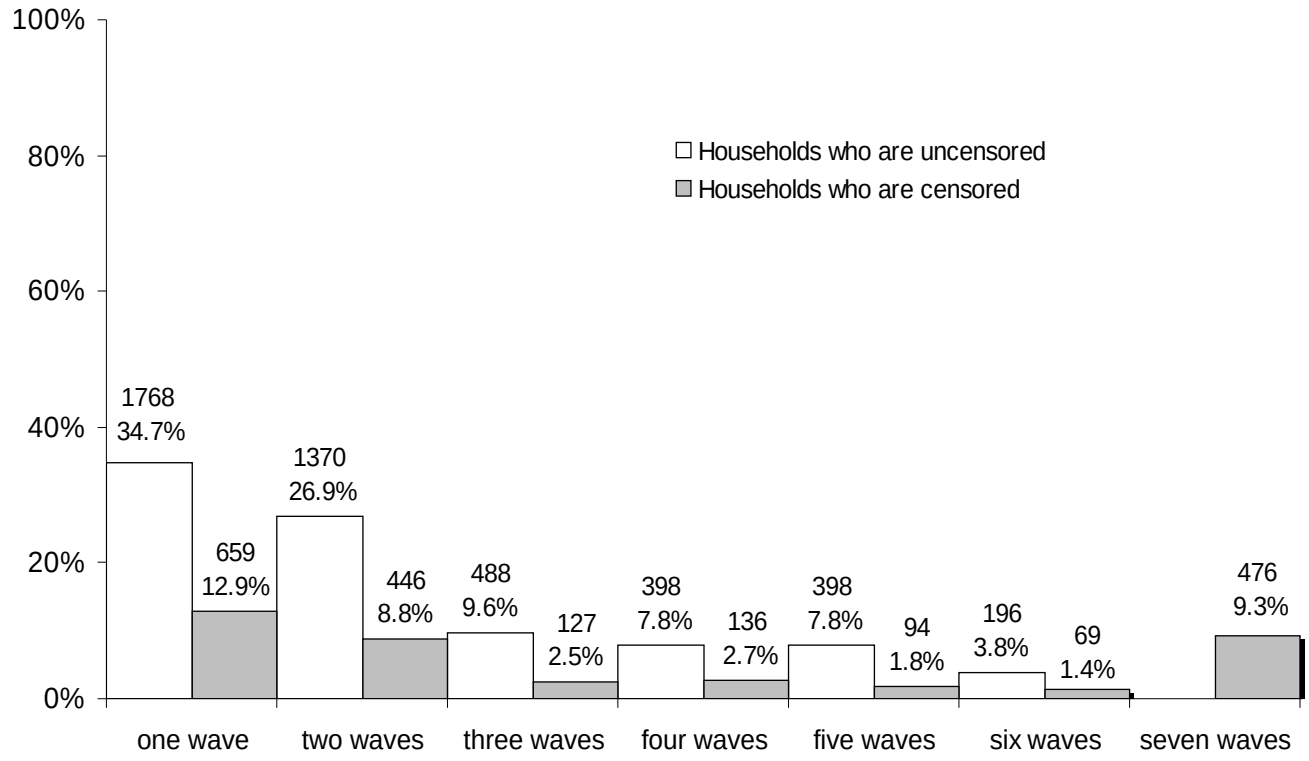

Figure 4-3: Households survey participation duration

4.3 IMPUTATION FOR ITEM NONRESPONSE

4.3.1 Imputation for Household Demographics

The PSTS data indicate that a total of 5094 households returned fully or partially completed survey diaries during the seven-wave survey. Each wave has three sets of data including household demographics, person demographics, and travel data. Because the household is the survey participant unit, our nonresponse study is based at the household level. Consequently, the household data are our primary focus and are assembled for the estimation of duration models. However, a certain number of households did not provide all the information on household demographics. For example, a total of 421 households (8%) filled out the travel diary without providing any household information at all. There are a larger number of households (1160, 23%) without any income information available for at least one wave. To avoid further reduction of the number of survey participants due to missing items, two methods are adopted to impute the missing household demographic information. One method is to develop models to impute missing variables. The other is to impute the missing item using other data sources, such as the person demographic data or the household data from the previous wave.

In the household demographic file, some households did not provide the information of household income and the number of household vehicles. Three types of income information were collected throughout the panel surveys. One is the exact household income in dollars. The second is a classification variable indicating whether the household income is under or beyond $35K. The third is a

categorical variable which divides income into eight classes. Compared to the categorical income variable, far more households did not provide the exact household income. The percentage of households with exact income missing ranges from 36.4% to 56.7% from wave to wave. The large number of households with missing income makes us cautious about imputation using linear regression model, because of the poor coverage of the remaining sample. On the other hand, most of the survey participants did provide the categorical income information. The households with missing categorical income comprise less than 10% of the survey participants for each wave. The available information on the income classes makes the imputation on the categorical income variable more reliable. Therefore, our imputation efforts are on the third categorical income variable. In this study an ordered response choice model is adopted to impute income (Bhat, 1994). The same model structure is also applied to impute the number of vehicles owned by households.

Other household variables, such as household size and household type, also have a small number of missing cells for some waves. These variables are imputed as follows. If the variables in the previous and the following wave are not missing and they are of the same value, the missing cell is imputed by the value of the variable in the previous wave. If the variables in the previous and the following wave are not of the same value, either the value of the previous wave or that of the following wave is randomly chosen to impute the variable in the missing wave. Among 421 households which do not have any household demographic information provided, 340 households' missing demographic

variables (other than income and the number of vehicles) are imputed by the values of the previous or the following wave.

There are a total of 41 households that only participated in one wave of the survey. Neither the previous nor the following wave data were recorded. The missing variables for these households are imputed by the information from the person data. The person data contains personal demographics of household members who are older than 15. Based on the variables in the person data, the household demographics can be obtained. One concern for using the person data to impute the household variables is that the person data files do not have information on children who are younger than 15, which may lead to misspecification of household type and underestimation of household size. However, 41 out of these 44 households are split households. Both the household data and the person data indicate that the split households were generated by young adults moving out of their parents' house. The data also shows that many of the split households have only one member. This suggests that the split households are less likely to have any children younger than 15. Therefore, the household demographics can be acquired by the aggregation of the person data without loss of accuracy.

The rest of the 30 households returned travel diary for one wave without providing any information for the household and person demographics. After the imputation procedure these 30 households still have most of the household variables missing and are removed from the study.

4.3.2 Imputation for Trip Characteristics

Just as some participating households returned travel diaries with no household demographic information provided, some others simply answered questions on household demographics but did not fill out the travel diary. Throughout the seven-wave survey, a total of 734 households had at least one wave of the travel data missing with household demographic information available. Among these households, 60.8% (446 households) did not return the travel diary for wave 6 and 38.8% of them (285 households) only participated for one wave. It is difficult to further investigate why so many households did not return the travel diary for wave 6. One possible reason is that the data were collected from May to August for wave 6, while the other surveys were conducted during the period of September to February. Many families take vacations in summer, so they may decline the survey request. In addition, the variation of household's travel pattern across seasons may have some impact on the survey participation. The famous September-to-April rainy season in the Seattle area may intensify the effects. Later in our model specification, we will accommodate the seasonal impact on the survey nonresponse.

Among 734 households with no travel information, only one household had travel data missing for two consecutive waves. All the other households did not return the travel diary for one wave and then stopped responding from the following wave survey completely. Again, the participation in the survey without providing all the necessary information is an indicator of quitting the survey for the next wave.

The number of households with travel data missing for each wave is shown in Table 4-2. In the table, the number of households is further stratified by survey participation duration. Missing travel data happens in four out of seven waves: wave 2, wave 3, wave 4, and wave 6. Similarly, no clear pattern can be observed to indicate why some households did not fill out the travel diary. One observation is that, in wave 6, a large number of refreshment households (277 out of a total of 946 newly recruited survey participants) did not provide any travel information and stopped responding thereafter. The high rate of travel data missing for refreshment samples may be closely related to the recruiting method and/or the data collection procedure for this particular wave. However, the exact guidelines used to recruit refreshment samples for wave 6 are not available and the information on data collection efforts is limited. Therefore, it is not reliable to draw any conclusions on the cause of nonresponse without more information and further analysis.

Table 4-2: Survey participation of households with travel data missing

Wave	Duration (waves)							Total
	1	2	3	4	5	6	7	
Wave two	3	117	1*	-	-	-	-	121
Wave three	-	29	68*	-	-	-	-	97
Wave four	5	29	3	34		-	-	71
Wave six	277	55	43	16	10	45	-	446
Total	285	230	115	50	10	45	-	735

Since we are particularly interested in the relationship of trip frequencies and survey participation, the 14% (734) of the households with missing trip characteristics is critical for the analysis. If we do nothing and simply remove these households from the analysis, the result would be a significant reduction of household observations and the loss of information since many removed households did supply travel information for the previous waves. Thus, the travel information is imputed for these households.

The travel data consists of as many as 50 variables describing when, where, and with whom the travel activity took place. It is challenging to impute each one of these variables reasonably. As trip frequency is our primary interest, we focus on imputing the frequencies of home-based work, home-based non-work, and non-home-based trips. The trip frequencies are imputed using the travel data of the previous wave. After the imputation, there are still a total of 285 households who only partially participated for one wave with trip frequencies

* One household had travel data missing for two waves (wave two and wave three).

missing. They comprise about 10% of the households with one wave duration and most of them entered the survey at wave 6. Because no other information on trip frequencies is available, these households are removed from the study.

Finally, after data cleaning and imputation, the data set prepared for the panel attrition analysis includes a total of 4802 households. The data contains not only household demographic variables, but also information related to the survey, such as which sample group a household belongs to and how many household members filled out the travel diary, as well as the trip frequencies of various trip purposes.

4.4 TRENDS IN HOUSEHOLD DEMOGRAPHICS

4.4.1 Household Location and Household Type

The households entering the first wave of the PSTP resided in one of the four counties (King, Kitsap, Pierce, and Snohomish) of the Central Puget Sound metropolitan region. Besides travel mode, the survey sample is also stratified by the county of residence (Murakami and Watterson, 1991). Furthermore, the refreshment samples were selected in the following waves with the criteria of replicating the sample strata of the first wave (including travel mode and household location), as well as the household life cycle. Because of the sample refreshment strategy, the sample fractions of residential location and household type across waves should be at a comparable level. The number of households by county of residence and wave of the survey did not vary much, as shown in Table 4-3. The percentage indicates the fraction of sample households that resided in

each county for each wave. The majority of the survey participants were living in King County. The percentages range from 41% to 47.4% across waves. About a tenth of the households were from Kitsap County. Approximately one quarter of the households resided in Pierce County and another quarter from Snohomish County.

The frequency and sample fraction of households by county of residence and the duration of survey participation are shown in Table 4-4a and Table 4-4b. Table 4-4a includes all households who did not change their household location while Table 4-4b demonstrates the participation duration for the households who moved during the survey period.

Table 4-3: Households by residence location and wave of the survey

County of Residence	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 4**	**Wave 5**	**Wave 6**	**Wave 7**
King	707 41.3%	825 41.0%	800 42.6%	927 44.5%	857 44.2%	916 46.7%	952 47.4%
Kitsap	204 11.9%	259 12.9%	220 11.7%	226 10.8%	209 10.8%	195 9.9%	213 10.6%
Pierce	366 21.4%	411 20.4%	367 19.6%	387 18.6%	379 19.6%	386 19.7%	435 21.7%
Snohomish	434 25.4%	517 25.7%	489 26.1%	542 26.0%	491 25.3%	459 23.4%	406 20.2%
Other counties	1 0.1%	1 0.1%	-	1 0.1%	2 0.1%	5 0.3%	1 0.1%
Total	1712	2013	1876	2083	1938	1961	2007

A total of 4149 households did not move during the survey period. As panel surveys continued, the participating households declined. The descriptive analysis indicates that households from Kitsap County and Snohomish County have slightly higher shares for longer participation (duration of four or more waves). Nevertheless, there is no clear trend indicating that households in various geographic locations can be differentiated much in terms of participation duration in the panel surveys.

Table 4-4a: Households by residence location and participation duration

County	One Wave	Two Waves	Three Waves	Four Waves	Five Waves	Six Waves	Seven Waves	Total
King	752 39.2%	520 27.1%	205 10.7%	121 6.3%	137 7.1%	56 2.9%	129 6.7%	1920
Kitsap	141 34.7%	95 23.4%	34 8.4%	30 7.4%	34 8.4%	24 5.9%	48 11.8%	406
Pierce	358 40.6%	232 26.3%	75 8.5%	48 5.4%	55 6.2%	28 3.2%	85 9.6%	881
Snohomish	289 30.8%	260 27.7%	85 9.1%	79 8.4%	90 9.6%	37 3.9%	98 10.4%	938
Other counties	2 50%	2 50%	-	-	-	-	-	4

During the survey period, a total of 653 households moved at least once and still managed to respond to the survey. The change of household location has an enormous impact on mail survey. One reason is that the relocation of the household is often associated with life cycle change, such as college graduation or

marital status. It may result in an abrupt shift in the opinion towards the survey and therefore lead to terminating the participation. Furthermore, the challenge for the data collection effort is to track the moved households. The commonly believed hypothesis is that relocated households have a lower response rate than the other sample segments. However, this hypothesis is not necessarily equivalent to the behavioral assumption that they are less likely to respond than the others without further accommodating the tracking problem. Among the PSTP participants, some households relocated within a county, some moved across counties, and some others did both. 81% of relocated households moved within the same county, 17% moved at least twice, and 11 households moved across counties, as shown in Table 4-4b. Similar to those who did not move, most of these relocated households stayed in the panel for a short period (less than three waves). The descriptive statistics do not show that the household location change caused a lower response rate in this case. However, it should be pointed out that the statement is conditioned on the initial condition, i.e., among households who responded to the survey for at least one wave.

Table 4-4b: Participation duration for relocated households

	Two waves	**Three waves**	**Four waves**	**Five waves**	**Six waves**	**Seven waves**	**Total**
Moved within a county	195 36.5%	69 13.0%	79 14.8%	57 10.7%	41 7.7%	91 17.1%	532
Moved across counties	5 45.5%	3 27.3%			2 18.2%	1 9.1%	11
Moved within a county and across counties	33 30.0%	24 21.8%	11 10.0%	18 16.4%	8 7.3%	16 14.5%	110

The household life cycle category is defined on the basis of household structure and household members' ages. The eight categories are:

- Households with any child under the age of 5;
- Households with all children between the age of 6 to 17;
- Households with no children and one adult at the age of 18 to 35;
- Households with no children and one adult at the age of 36 to 64;
- Households with no children and one adult older than 65;
- Households with no children and two or more adults at the age of 18 to 35;
- Households with no children and two or more adults at the age of 36 to 64;
- Households with no children and two or more adults older than 65.

The sample distribution of the household types is shown in Table 4-5 for each wave. Due to the panel attrition and refreshment samples, the sample size varies across waves. Wave 1 contains a total of 1712 households and wave 4 has the highest sample size of 2083. Therefore, the direct comparison of household

frequency in each life cycle category may not be as informative as evaluating the percentage change. The data show that the sample fraction of households with any child under five years old decreases from 16.9% to 11.9%, while the sample shares of aging households, especially those with adults older than 65, increase throughout the panel survey. The sample fractions for other household types remain at a steady level. The variation across waves is controlled within 2%.

Table 4-5: Households by life cycle type and wave

Household type	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7
Any child under age 5	290 16.9%	333 16.5%	256 13.6%	293 14.1%	249 12.8%	253 12.9%	238 11.9%
All children between age 6 to 17	336 19.6%	420 20.9%	401 21.4%	425 20.4%	389 20.1%	406 20.7%	372 18.5%
One adult under age 35, no children	78 4.6%	93 4.6%	65 3.5%	109 5.2%	59 3.0%	49 2.5%	62 3.1%
One adult between age 36 to 64, no children	155 9.1%	188 9.3%	194 10.3%	205 9.8%	213 11.0%	209 10.7%	226 11.3%
One adult older than 65, no children	73 4.3%	88 4.4%	100 5.3%	106 5.1%	115 5.9%	123 6.3%	166 8.3%
Two+ adults under age 35, no children	116 6.8%	140 7.0%	107 5.7%	161 7.7%	106 5.5%	105 5.4%	104 5.2%
Two+ adults between age 36 to 64, no children	474 27.7%	518 25.7%	493 26.3%	521 25.0%	534 27.6%	520 26.5%	574 28.6%
Two+ adults older than 65, no children	190 11.1%	233 11.6%	260 13.9%	263 12.6%	273 14.1%	296 15.1%	265 13.2%
Total	1712	2013	1876	2083	1938	1961	2007

The decrease of households with small children and the increase of households with older adults may reflect their willingness for continuing participation in the panel survey. Table 4-6 compares the frequency and percentage of households in each life cycle type by the duration of participation. There are many more households participating in the survey for one or two waves than for three or more waves. An abrupt decline in households is observed between two-wave and three-wave participation. However, the phenomenon is not particularly correlated with any particular household life cycle type. Thus, we cannot draw any solid conclusion about the participation pattern in the panel surveys with regard to the household type based on the descriptive statistics, except that the sample fraction of the households with one and only adult under age 35 consistently decreases while the participation duration increases.

Table 4-6: Households by life cycle type and survey participation duration

Household type	One wave	Two waves	Three waves	Four waves	Five waves	Six waves	Seven waves
Any child under age 5	266 17.3%	227 16.9%	84 17.0%	69 18.8%	55 14.1%	40 20.4%	59 12.6%
All children between age 6 to 17	266 17.3%	268 20.0%	91 18.4%	77 20.9%	85 21.7%	47 24.0%	79 16.9%
One adult under age 35, no children	139 9.0%	98 7.3%	24 4.8%	15 4.1%	10 2.6%	4 2.0%	9 1.9%
One adult between age 36 to 64, no children	159 10.3%	109 8.1%	59 11.9%	46 12.5%	37 9.5%	15 7.7%	48 10.3%
One adult older than 65, no children	83 5.4%	62 4.6%	20 4.0%	13 3.5%	26 6.6%	6 3.1%	19 4.1%
Two+ adults under age 35, no children	203 13.2%	150 11.2%	50 10.1%	26 7.1%	17 4.3%	14 7.1%	15 3.2%
Two+ adults between 36 to 64, no children	316 20.5%	310 23.1%	124 25.1%	82 22.3%	116 29.7%	40 20.4%	168 35.9%
Two+ adults older than 65, no children	110 7.1%	118 8.8%	43 8.7%	40 10.9%	45 11.5%	30 15.3%	71 15.2%
Total	1542	1342	495	368	391	196	468

4.4.2 Household Income and Vehicle Ownership

Income and vehicle ownership are two major economic indicators for households. The number of households segmented by income category and vehicle ownership is shown in Table 4-7 and Table 4-8, respectively.

Table 4-7: Households by income category

Income Category	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7
<10K	35 2.0%	39 1.9%	41 2.2%	89 4.3%	100 5.2%	43 2.2%	136 6.8%
10K – 15K	133 7.8%	124 6.2%	69 3.7%	75 3.6%	60 3.1%	62 3.2%	51 2.5%
15K – 25K	278 16.2%	294 14.6%	271 14.4%	257 12.3%	188 9.7%	162 8.3%	161 8.0%
25K – 35K	205 12.0%	162 8.0%	408 21.7%	383 18.4%	351 18.1%	286 14.6%	280 14.0%
35K – 45K	262 15.3%	274 13.6%	341 18.2%	400 19.2%	367 18.9%	419 21.4%	325 16.2%
45K – 55K	474 27.7%	527 26.2%	289 15.4%	311 14.9%	318 16.4%	312 15.9%	318 15.8%
55K-75K	209 12.2%	380 18.9%	270 14.4%	323 15.5%	303 15.6%	340 17.3%	358 17.8%
>75K	116 6.8%	213 10.6%	187 10.0%	245 11.8%	251 13.0%	337 17.2%	378 18.8%
Total	1712	2013	1876	2083	1938	1961	2007

Table 4-8: Households by vehicle ownership

Number of vehicles	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7
0	64 3.7%	52 2.6%	38 2.0%	73 3.5%	70 3.6%	70 3.6%	66 3.3%
1	417 24.4%	500 24.8%	434 23.1%	537 25.8%	472 24.4%	474 24.2%	520 25.9%
2	740 43.2%	875 43.5%	865 46.1%	922 44.3%	864 44.6%	895 45.6%	869 43.3%
3	321 18.8%	410 20.4%	345 18.4%	375 18.0%	373 19.2%	359 18.3%	382 19.0%
4	112 6.5%	109 5.4%	125 6.7%	127 6.1%	106 5.5%	105 5.4%	102 5.1%
>= 5	58 3.4%	67 3.3%	69 3.7%	49 2.4%	53 2.7%	58 3.0%	67 3.4%
Total	1712	2013	1876	2083	1938	1961	2007

The sample fraction of the eight income categories fluctuates erratically from wave to wave. The fluctuation may be due to the following reasons. First, income of the population changed even as the panel data were collected. According to *Puget Sound Trends*, King County per capita income estimates have increased more rapidly than the rest of Washington State since 1995. Because many software companies in the Puget Sound region are located in King County, and the booming era of IT industry in mid-90s attracted many professionals moving into the region, it is reasonable that more households fell into the high income category (income > 55K) for wave 5 through wave 7. Second, the small

sample size may, to some extent, contribute to the abrupt fluctuation. The estimated number of households in the region is 1,269,070 in 1999. The PSTP sampled only about 0.5% of the whole population. Also, measurement error for the income variables is probably greater than other variables. Therefore, it is not surprising to observe some variation in the panel data.

If we segment the households by the number of household owned vehicles, the sample fractions remain steady across waves. Vehicle ownership is often strongly correlated with income. However, it is also found that income generally has a lagged impact on the vehicle ownership. This lagged impact may explain why simultaneous fluctuations are not observed in both income and vehicle segmentations.

4.4.3 Household Size and Workers

The household distribution by household size is shown in Table 4-9. The sample data show that the number of single-member households goes up along the panel survey. The phenomenon could be a reflection of the generation of the split households from their mother households who already participated in the survey, or an indicator of increase of single-member households in the population. Most of the split households are single-member households. The number of split households in each wave is shown in Table 4-10. Wave 2 (with 42 split households) and wave 4 (with 33 split households) have more split households than other waves. Few split households appear in wave 3, wave 5, and wave 7.

Despite the existence of the split households, the sample still demonstrates a pattern of increase of single-member households.

The sample fraction of households with more members varies slightly from wave to wave. The variations are within 3% across waves with no clear pattern of increase or decrease. The sample fraction remains relatively stable for these households.

Table 4-9: Households by household size

Household size	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 4**	**Wave 5**	**Wave 6**	**Wave 7**
1	306 17.9%	366 18.2%	359 19.1%	421 20.2%	387 20.0%	380 19.4%	454 22.6%
2	693 40.5%	794 39.4%	789 42.1%	851 40.9%	829 42.8%	795 40.5%	817 40.7%
3	294 17.2%	357 17.7%	292 15.6%	338 16.2%	280 14.4%	321 16.4%	304 15.1%
4	287 16.8%	351 17.4%	316 16.8%	326 15.7%	309 15.9%	319 16.3%	285 14.2%
5	93 5.4%	95 4.7%	89 4.7%	99 4.8%	94 4.9%	103 5.3%	109 5.4%
>= 6	39 2.3%	50 2.5%	31 1.7%	48 2.3%	39 2.0%	43 2.2%	38 1.9%
Total	1712	2013	1876	2083	1938	1961	2007

Table 4-10: Number of split households by wave

Wave	Split households	Total households
Wave 1	0	1712
Wave 2	42	2013
Wave 3	6	1876
Wave 4	33	2083
Wave 5	2	1938
Wave 6	0	1961
Wave 7	14	2007

The household distribution by the number of workers is shown in Table 4-11. In the sample, the shares of one-worker and two-worker households remain about 35% from wave 1 to wave 7. However, the zero-worker households jumped from 17.4% in wave 1 to 30% in wave 6. The sample fraction of three-plus-worker households is at the highest level of 6.5% in wave 1 and then down to around 3% for the rest of the waves.

Table 4-11: Households by the number of workers

Household workers	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7
0	298 17.4%	560 27.8%	552 29.4%	555 26.6%	564 29.1%	588 30.0%	506 25.2%
1	660 38.6%	700 34.8%	577 30.8%	802 38.5%	718 37.0%	634 32.3%	761 37.9%
2	643 37.6%	691 34.3%	689 36.7%	670 32.2%	615 31.7%	684 34.9%	672 33.5%
>= 3	111 6.5%	62 3.1%	58 3.1%	56 2.7%	41 2.1%	55 2.8%	68 3.4%
Total	1712	2013	1876	2083	1938	1961	2007

4.5 TRENDS IN TRAVEL ACTIVITIES

The travel data summarize the information in the travel diaries maintained by eligible household members (15+ years old) over two consecutive weekdays. The variables describing the trip making characteristics belong to each of the following categories: trip purpose, travel mode, travel partner, departure and arrival time, and origin and destination of the travel. As trip frequency can be used as an indicator of survey burden for travel activity surveys and the sampling strata is based on travel mode, the descriptive analysis focuses on the trip rates by trip purpose and travel mode.

In the PSTP survey, sample units are selected from regular driving-alone, carpooling, and transit-user households. The number of households by sampling

group and survey participation duration is shown in Table 4-12. When the sample households are segmented by the survey participation duration, the fractions of regular driving-alone households remain rather stable. Among households participating only for one wave, 65% are regular driving-alone households, while among households participating in all seven waves of the survey, 65.8% are regular driving-alone households. A similar relatively stable pattern is observed for regular carpooling households too. For regular transit-user households, the fractions slightly decrease for longer survey participation duration. 14.2% of the households with one-wave duration are regular transit-user households and only 10.7% for the households with seven-wave duration.

Table 4-12: Households by sampling group and survey participation duration

	One wave	**Two waves**	**Three waves**	**Four waves**	**Five waves**	**Six waves**	**Seven waves**
Regular driving-alone households	1003 65.0%	849 63.3%	317 64.0%	249 67.7%	273 69.8%	129 65.8%	308 65.8%
Regular carpooling households	322 20.9%	320 23.8%	119 24.0%	74 20.1%	75 19.2%	42 21.4%	110 23.5%
Regular transit-user households	217 14.2%	173 12.9%	59 11.9%	45 12.2%	43 11.0%	25 12.8%	50 10.7%
Total	1542	1342	495	368	391	196	468

The distributions of trip rates are shown in Figure 4-4 to Figure 4-9. As we expected, the distributions of trip rates zigzag with peaks at even numbers no matter whether the number of trips are segmented by trip purpose or travel mode. The trip rate distributions are similar for all waves. The diagrams show that many

of the households choose to drive alone while undertaking trips. Few households use a ridesharing mode and the majority of the households (about 1300 to 1600) never use transit.

Another pattern we notice is that the zero-trip households tend to increase in the later waves. For home-based work trips, 359 (20.9%) households made no trips in wave 1 and this number goes up to 475 (23.7%) for wave 7. A similar pattern is also observed in home-based non-work, non-home-based, driving-alone, and carpooling trips. The increasing number of zero-trip households may suggest a self-selection pattern of households. Except for an increasing number of zero-trip households, no other clear trends are observed when comparing the trip rates across waves.

4.6 SUMMARY

The Puget Sound Transportation Panel data are used for the empirical analysis in this dissertation. This chapter first gives a brief background introduction on the data collection process, and then it describes the data cleaning procedure to prepare data for the model estimation. Item nonresponses in the data are imputed in various methods to avoid further reductions in sample size. Some missing variables are imputed using information from different data sources in different waves. Other missing variables are imputed by statistical models.

The data descriptions are focused on household demographics and travel characteristics, segmented by either survey participation duration or the time of the survey. The descriptive statistics generally cannot provide convincing

evidences on survey participation behavior. When the data are segmented by the number of waves for which a household has participated in, the descriptive statistics often reveal a steady sample composition across all the segments due to the limitation of descriptive statistics in multivariate analysis. The next chapter presents an econometric model to examine households' participation duration in the panel surveys. The strength of the model lies in its ability to isolate the marginal effects of different explanatory variables on the dependent variable of interest.

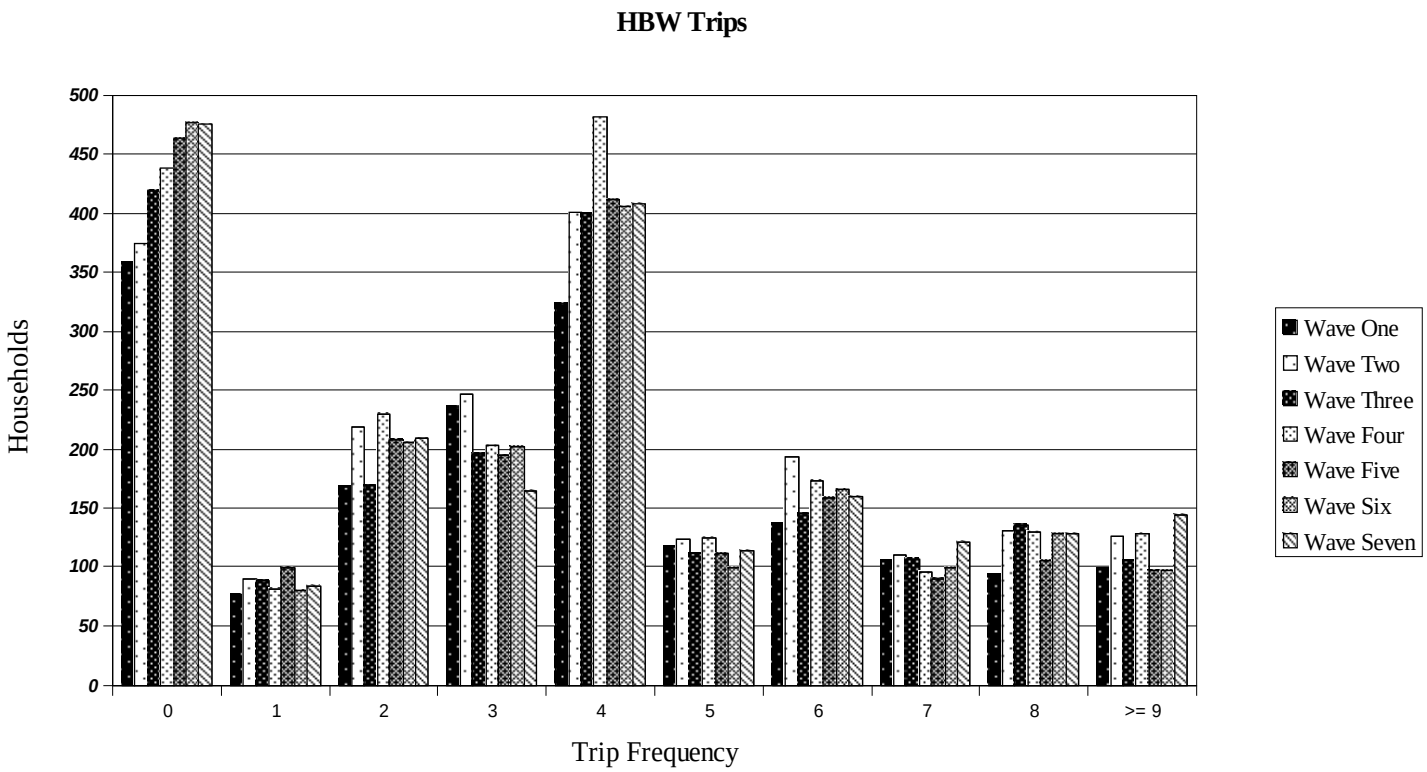

Figure 4-4: Trip frequency for home-based work trips

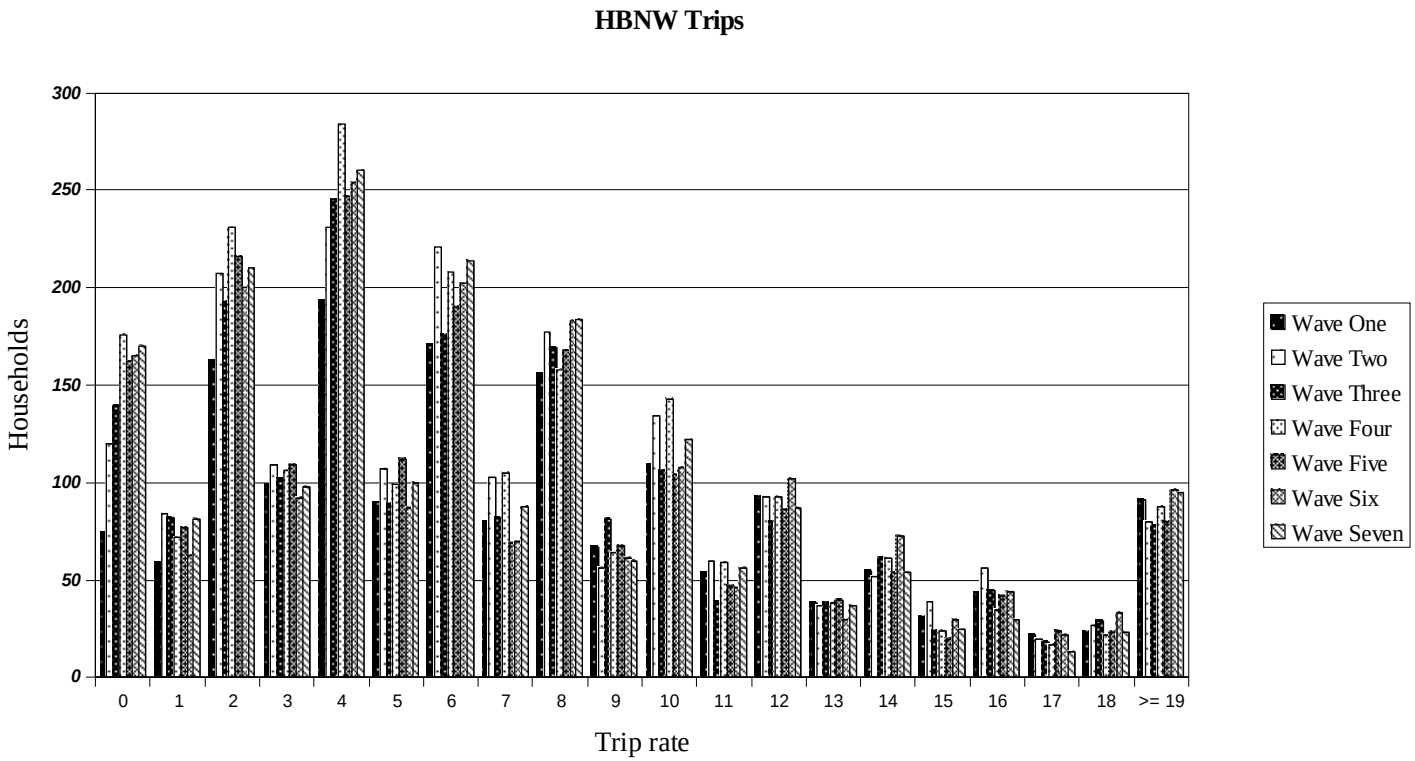

Figure 4-5: Trip frequency for home-based non-work trips

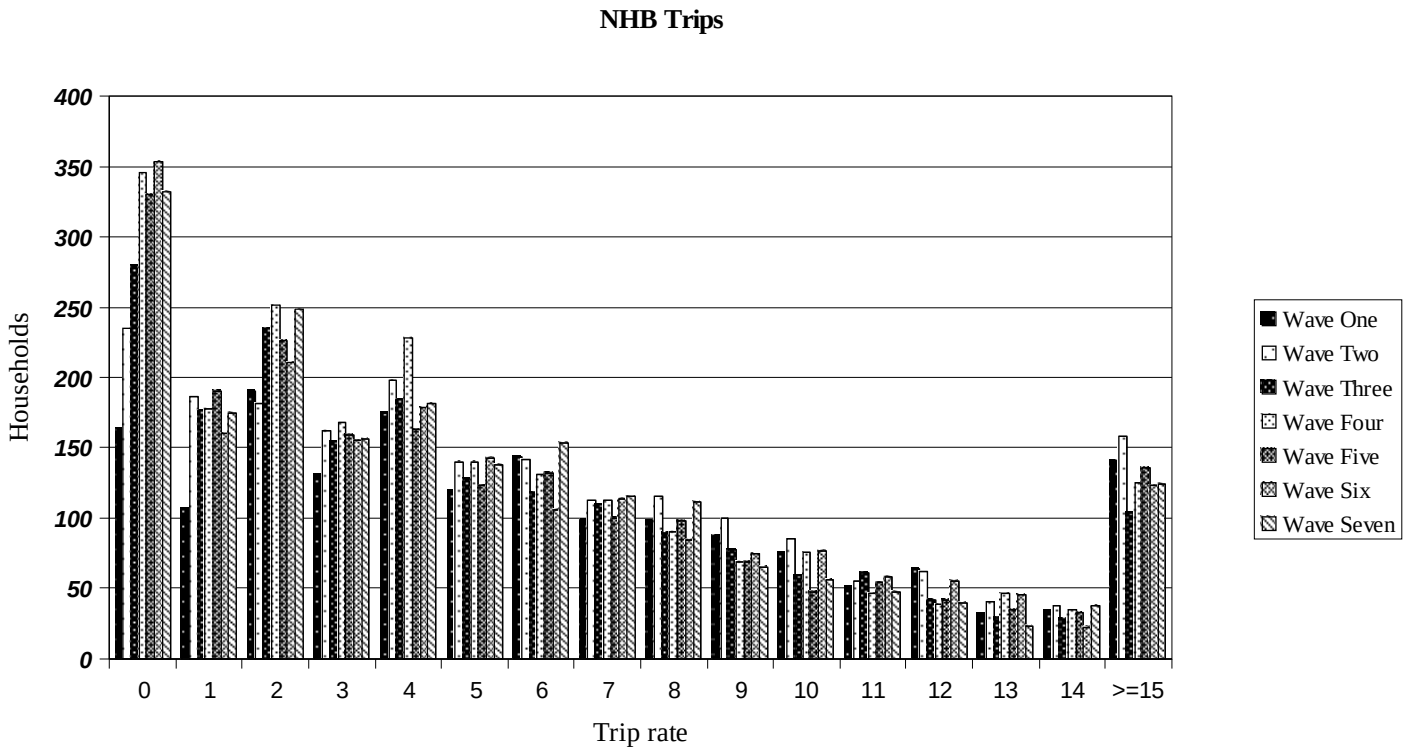

Figure 4-6: Trip frequency for non-home-based trips

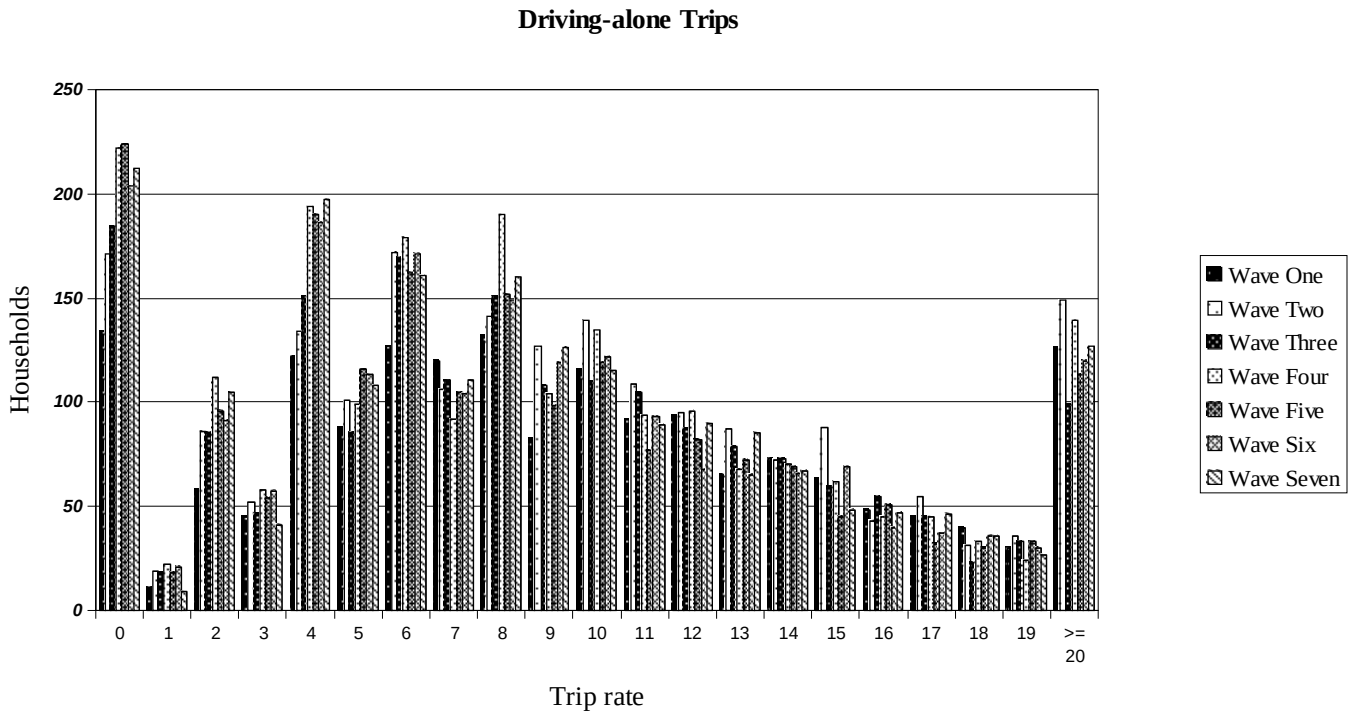

Figure 4-7: Trip frequency for driving-along trips

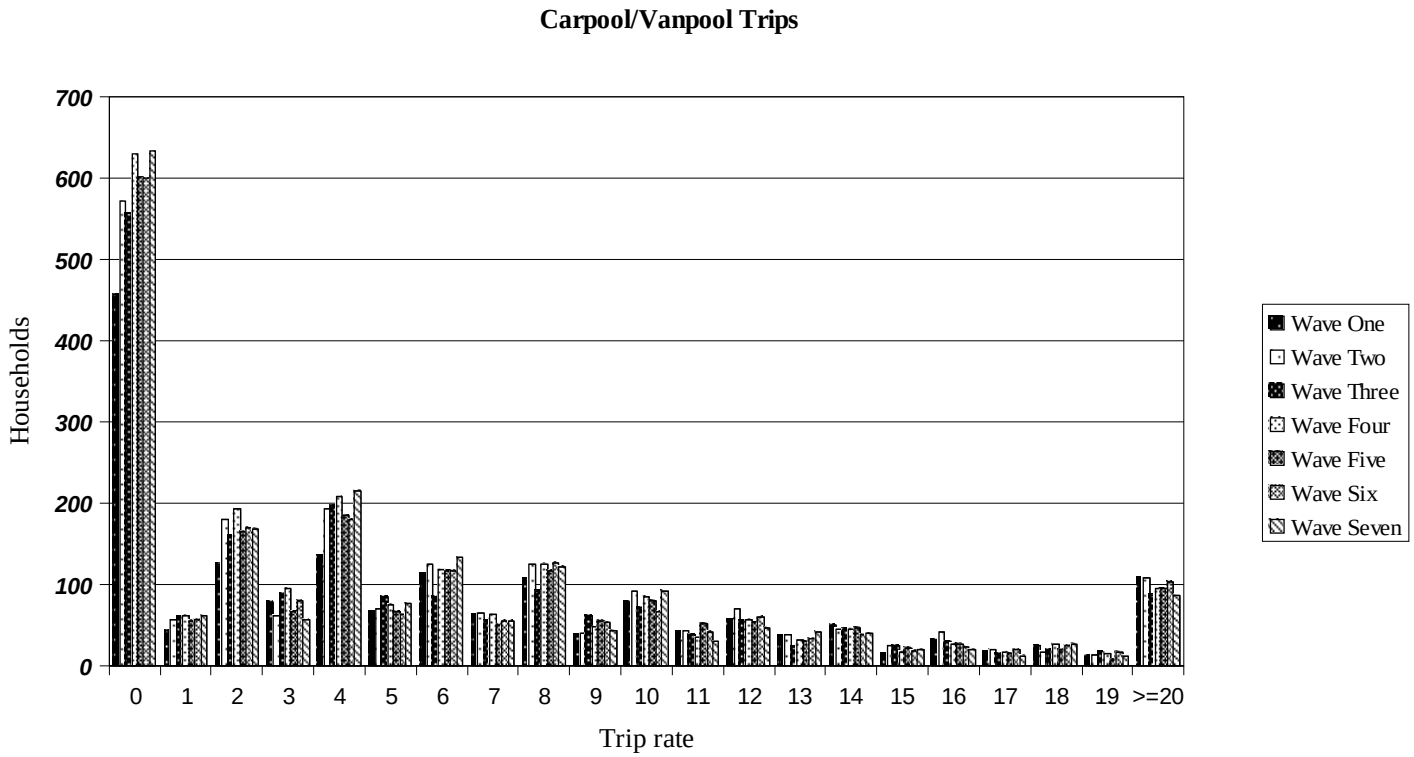

Figure 4-8: Trip frequency for carpool/vanpool trips

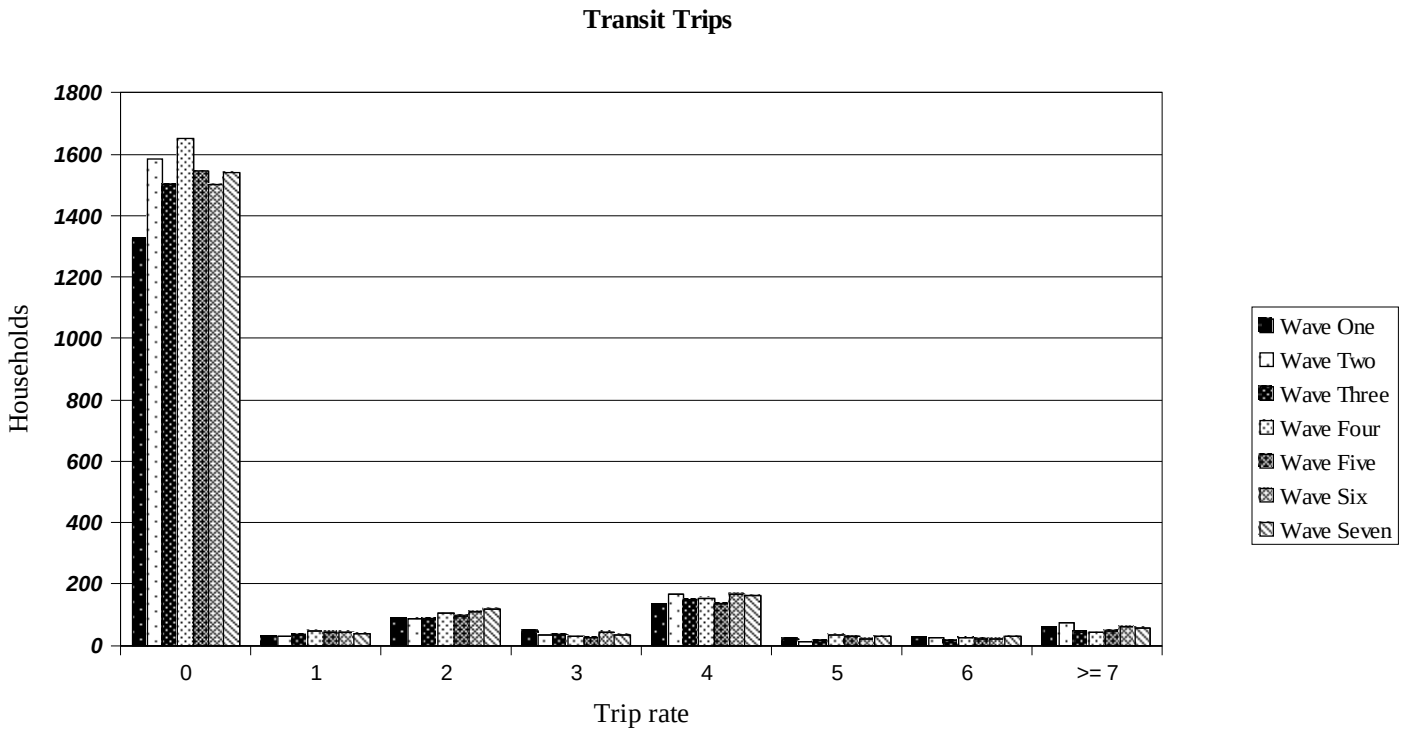

Figure 4-9: Trip rates by travel mode (transit trips)

Chapter 5 Analysis of Survey Participation Duration with Trip Frequencies as Exogenous Variables

This chapter proposes a hazard-based duration model for the participation duration in panel surveys. Section 5.1 gives a general introduction to the duration model, especially on proportional hazard models. The properties of the models, such as distribution of hazard function, covariate effects, and heterogeneity are also discussed in this section. Section 5.2 describes the model structure adopted to model the number of waves in which households have continuously participated in panel surveys. The base model is developed without consideration of the unobserved heterogeneity while the second model accommodates heterogeneity by a disturbance term. Time-varying covariates are implemented in both models. The empirical results based on the PSTP are presented in Section 5.3. Since binary logit model is a common approach to analyze the survey participation decision, Section 5.4 evaluates these two methods via comparisons of covariate effects and model prediction.

5.1 DURATION MODEL

5.1.1 Introduction

Duration models were initially proposed to analyze duration data (or failure time data), such as survival time of potential heart transplant recipients and problems arising in the fields of biomedical science and industrial engineering. The models have been widely used in many scientific disciplines, including

economics (Lancaster, 1979; Han and Hausman, 1990), marketing research (Vilcassim and Jain, 1991), and transportation studies (Bhat, 1996). The focus of duration models is the use of the hazard function, i.e., the end-of-duration occurrence conditional upon the fact that the duration has lasted to some specific time (Kiefer, 1988; Hensher and Mannering, 1994).

There are conceptual and intuitive reasons to consider the conditional probability instead of the unconditional probability when analyzing duration data, although Kiefer (1988) indicated an exact mathematic equivalence between the hazard function and the unconditional probability of the occurrence of an event. In reality, we make many decisions that follow certain sequences. From a behavioral standpoint, each decision-making procedure is conditional on what have happened prior to this particular event. This conditional property is more obvious when modeling a decision to terminate an activity. The activity can only be terminated given that it has lasted to the point of termination. Thus, the variable of interest in duration models intuitively fits the behavior process. Furthermore, the conditional probability accommodates the dynamics of duration explicitly. Consider the duration process as a finite number of sequential trials for the discrete cases, and the discrete time cases can be generalized to the continuous time cases when the time interval is infinitely small. Every trial could be independent of each other. This assumption indicates the constant hazard, i.e., the likelihood of failure at time *t*, conditional upon duration up to time *t*, is constant over time. Or, the occurrence of a trial could be positively (negatively) associated with the duration for which the event has not occurred. The longer the duration

is, the more (less) likely it is that the event would occur. This assumption implies the positive (negative) duration dependence. Thus, the hazard function is convenient for behavioral interpretation and hypothesis testing.

Besides conceptual and intuitive reasons, the duration models also have methodological advantages over traditional method in modeling survival time. The models overcome the problems of accounting for censored observations and time-varying explanatory variables that arise when applying standard regression models to duration data. The common cause of censoring in duration data is that the measurement is made while the process is ongoing, or the process is continuing beyond the period of measurement. The duration of a censored observation is at least the observed time period t_i, but not equal to it. The estimation procedure of the conventional regression model, which is based on fixed values of the dependent variable, can not account for the censored nature of duration data. In regression analysis, when the dependent variable is censored, one common treatment is to transform the values in a certain range to a single value. Not surprisingly, this ignorance of censoring loses some information of data and may lead to inconsistent estimates. In addition, when measuring duration, the observations have been underway for a period of time. Therefore, the observed covariates may have changed during this time interval. It is considerably more complicated to incorporate time-varying covariates in regression models, while hazard-based duration models can accommodate them in a relatively simple manner. The incorporation of time-varying covariates in the model will be discussed in the next section.

In this dissertation, we propose a hazard-based duration model to analyze the attrition behavior in multi-wave panel surveys. The variable of interest is how many waves a household would stay in the panel. When the participation procedure is considered as a duration process, a household who has continuously been in the panel can be viewed as a 'survivor' of the panel. So the hazard-based duration model is applicable to the study. Additionally, one intuitive hypothesis of the analysis is that the response behavior of the survey participants relies upon their experience in the previous waves, especially upon the number of waves the households have participated in, i.e., the presence of duration dependence. The duration models have an advantage of capturing this duration dependence. Therefore, theoretically, the hazard-based duration model is plausible for the analysis.

Censoring and time-varying variables, the two distinguishing features of duration data, also characterize the response duration in the multi-wave panel surveys. Households could be right censored. Right censored households are those who remain in the last wave of the survey. Household demographic variables may change since the panel surveys have lasted for years. Meanwhile, the survey characteristics differ among the various waves. For example, the time intervals between two consecutive waves are not constant over time in PSTP, which may have some impact on the response rate. As previously mentioned in the literature review, one challenge to investigate the effectiveness of the survey method is the difficulty in differentiating the joint impact of various survey features. The proposed model structure provides an opportunity to overcome this

difficulty by allowing the characteristic variables of the survey feature as time-varying covariates in the model. The model results would present a significant insight into survey plan and survey management.

It is a common hypothesis that attrition behavior is correlated with the number of trips or activities undertaken during the survey period for household travel surveys. The probability of a household not responding to an activity survey, for example, may be higher if the household made a large number of trips during the survey period due to the extra effort required to report these trips, compared to a household that did not make any trips during the survey period. In this chapter, we consider the trip frequency as an exogenous variable in the hazard duration model of survey participation duration. Later we develop a model system considering the trip frequency and survey duration simultaneously. The model system will be discussed in Chapter 6.

5.1.2 Distribution of Hazard Function

The hazard can be expressed by a cumulative distribution function, $F(t)$, and a corresponding density function, $f(t)$. The cumulative probability of a non-negative random variable T can be specified as

$$F(t) = \Pr(T < t), \tag{5-1}$$

where Pr denotes the probability and t is a realization of T. T can be a continuous variable as well as a discrete variable. However, there is no significant difference in the methodology between a continuous T and a discrete T. The discrete T can be considered as a result of grouping the continuous time into several discrete

intervals. The corresponding density function is the first derivative of the cumulative distribution function with respect to time and can be written as

$$f(t) = dF(t)/dt . \tag{5-2}$$

When applying duration models, we are usually more interested in the probability that an event lasts longer than *t*. A corresponding term is survival function, *S*(*t*), also referred to as the endurance probability. The survival function is defined as

$$S(t) = 1 - F(t) = \Pr(T \geq t) . \tag{5-3}$$

The hazard rate $\lambda(t)$ is then defined as the instantaneous probability that the duration will end in a infinite small time period Δ, given that the duration has lasted until time *t*. The hazard rate can be written as

$$l(t) = \lim_{\Delta \to 0} \frac{\Pr(t \leq T \leq t + \Delta \mid T \geq t)}{\Delta} . \tag{5-4}$$

If the probability density function *f*(*t*) is continuous, the hazard function can be expressed as

$$l(t) = \frac{f(t)}{S(t)} = \frac{-d \ln S(t)}{dt} . \tag{5-5}$$

As mentioned in the previous section, the shape of the hazard function has important implications for duration dynamics. In many applications, the hazard function is clearly dependent on the length of time for which the duration process has lasted, which indicates the presence of the duration dependence. One example can be drawn from activity duration studies. The probability of an individual ending his shopping activity after the activity has lasted for sixty minutes may be higher than it was after the activity had lasted for twenty minutes.

If this is true, the hazard function slopes upward and the positive first derivative of the hazard function with respect to time would be observed, indicating the positive duration dependence. The opposite case is a decreasing hazard function or negative duration dependence. For example, consider the study of unemployment duration. A reasonable hypothesis is that the longer an individual is out of a job, the less competitive he/she is in the job market, and therefore the less likely it is that he/she will find a job and end unemployment in the next short time interval.

The hazard function can be modeled using either a parametric or a non-parametric form to analyze the duration dependence. In a parametric form, the duration distribution is pre-specified, and consequently, the hazard function is pre-determined. Generally, the distribution of duration process is chosen on the basis of computational convenience, a particular economic theory, or the preliminary analysis of the data. The form of the hazard function is then determined corresponding to the duration distribution assumption. The commonly used duration distributions are exponential, Weibull, log-normal, gamma, generalized gamma, and log-logistic distributions.

All these duration distributions are non-negative distributions. Among these distributions, the exponential and Weibull distributions are most commonly used. The exponential distribution is obtained by taking the hazard function to be constant, $\lambda(t) = \lambda$, over the range of T. With one parameter $\lambda > 0$, the exponential distribution imposes a fairly restricted assumption that excludes any duration dependence in the duration process. The Weibull distribution is a generalization

of the exponential distribution that allows the hazard to depend on the survival period. The two parameter Weibull distribution can be described by the following hazard function,

$$l(t) = lp(lt)^{p-1}, \qquad (5\text{-}6)$$

where $p > 0$. The hazard function is monotonically decreasing for $p < 1$, increasing for $P > 1$, and reduces to the constant hazard rate if $p = 1$. The Weibull hazard function provides a more flexible way to capture duration dependence. However, the monotonic property is in direct contradiction with the fact that the duration dependence itself may vary along the time horizon in some applications. For instance, in the economics study of strike duration, in the early stage when T is small, the negative duration dependence exists because the longer strike indicates the more severe problems that led to the strike in the first place. Therefore, it is less likely to end in the next short time interval. On the other hand, when the strike has lasted for a certain period of time, such as a critical threshold time period t^*, a positive duration dependence may be observed because the longer a strike persists, the more likely it is that the strike will end soon due to the increasing willingness of the involved parties to resolve the problem. The generalized gamma distribution may overcome the restriction of the monotonicity imposed on the hazard function. The distribution generalizes the gamma distribution by introducing a scale parameter. The hazard function corresponding to the generalized gamma distribution can be derived into many special cases, such as the exponential, gamma, Weibull, and log-normal distributions.

If a parametric hazard function is adopted, the shape of the duration distribution is pre-determined. The model estimation actually is to calibrate a few distribution parameters. However, the specification analysis after the model has been fit into a particular distribution may suggest that this pre-determined distribution is not adequate to describe the data. In this case, the non-parametric hazard function is another option with more degrees of freedom. It is especially suitable to adopt the non-parametric hazard form when little or no knowledge of the duration process is available. The non-parametric form must be based on the scale of discrete time, or cumulative time interval. Within each time interval the hazard rate is assumed to be a constant. Nonetheless, no other constraint is placed on the hazard shape. It should be noted that the number of time intervals determines the degrees of freedom of the non-parametric hazard model. Therefore, when transferring a continuous time scale into discrete time intervals, the length of the time interval should be carefully determined. In the application of panel survey participation duration, a natural choice of time interval would be the interval between the consecutive panel waves. The transformation of time scale is not a problem for this particular case. Given the fact that few hypotheses of duration dependence in panel survey participation have been tested in earlier literature, the non-parametric form seems more appropriate for this study.

In a parametric or a non-parametric form, the shape of the hazard function provides intuitive insights into the duration process. In addition, an individual's characteristics may play a significant role in the duration process and it is equally important to implement this effect in the model structure. The effect of these

individual demographic variables, or covariates, can be incorporated in the hazard-based duration model through two model structures: proportional hazard model and accelerated lifetime model. The difference between these two models is that the proportional hazard model assumes that the effects of covariates act multiplicatively on the baseline hazard, while the accelerated lifetime model assumes the covariates re-scale time directly in the survival function. In this study, we use the proportional hazard model structure to analyze the survey participation duration because the structure is flexible in modeling hazard rate and the covariates effect. The model is discussed in detail in the next section.

5.1.3 Proportional Hazard Duration Model

The proportional hazard duration model is widely used in many studies. In the model, the hazard function consists of two components. One is baseline hazard and the other is covariates effect. The proportional hazard function can be written as follows:

$$l(t, x, b, l_0) = l_0(t) f(x, b), \qquad (5\text{-}7)$$

where $\lambda_0(t)$ is the baseline hazard and $f(x, \beta)$ is the function accommodating the covariates effect. The typical specification of $f(x,\beta)$ is the exponential form. The exponential form is chosen because it guarantees the positive hazard rate without imposing any constraints on the estimated coefficient vector b. In this dissertation, the exponential form is used to capture the covariates effect. When the proportional hazard function is adopted to model a duration process, there are

three issues worthy of some considerations. These issues are: baseline hazard function, covariate effects, and individual heterogeneity.

5.1.3.1 Baseline Hazard Function

The baseline hazard function can take the parametric or non-parametric form. As discussed in the previous section, a parametric approach assumes that the duration distribution is mostly known with the exception that a few scalar parameters need to be estimated. The distribution assumption is generally chosen based on the preliminary knowledge of the data or simply for the computational convenience.

This study considers a household's response to the panel surveys as a duration process. It is found that few attrition studies have been performed on the multi-wave panels. In addition, it is often observed that nonresponse behavior is survey-specific. Therefore, any assumption of duration distribution of household's responding behavior would be arbitrary. In this sense, the non-parametric form has more flexibility and is more suitable for the study. Furthermore, even when the duration process actually follows a probability distribution while a non-parametric modeling approach is used, the estimates will still be consistent. The only deficiency is the loss of efficiency that may not be very substantial (Meyer, 1987). Consequently, a non-parametric baseline hazard is strongly recommended and we adopt this form for our model development (Bhat, 1996). With non-parametric baseline hazard function and an exponential

function accounting for the covariates effect, the proportional hazard model is referred to as semi-parametric proportional hazard model.

5.1.3.2 Covariates Effect

There are two options to incorporate covariates effect in the model. The first option is to use the covariates of initial wave, assuming that household demographics do not change over time. The other is to incorporate time-varying covariates in the model. It is important to include the time-varying covariates in the model because the household demographics do change over time. The demographic changes may affect households' attitudes toward survey participation. It would not be appropriate to estimate the model with only the household variables from the initial wave and to assume them unchanged. In addition, when households repeatedly participate in the survey, different households may experience different fatigue stage. Using time-varying covariates can reveal some scenarios of the fatigue stage for different household segments and therefore can improve the accuracy of the model. Thus, the time-varying covariates are implemented in the model structure.

Time-varying covariates can be classified into two broad categories: external and internal covariates (Kalbfleisch & Prentice, 1980). The external covariates are not directly involved with the failure mechanism, i.e. the covariates are determined in advance before the study. The internal covariates are observed only as long as the individual survives and is not censored. All these time-varying covariates are internal covariates.

The time-varying covariates can be incorporated in a straightforward manner (Bhat, 2000; Bhat and Steed, 2001). In our model, it is assumed that the household demographic variables observed in one wave will keep unchanged until the next wave. This assumption simplifies the incorporation of the time-varying covariates and the model estimation procedure. The covariates effect is then a cumulative function of time-varying variables along the entire time path.

5.1.3.3 Heterogeneity

The models discussed in the previous sections are based on the assumption that the duration distribution is homogenous over the population after controlling for the effect of explanatory household demographic variables. This assumption implies that the variation of the duration is fully captured by these explanatory variables. However, in reality often not every factor that might have impacts on the duration process is observed. When some variables do affect the duration process but are not observed and included in the model, the deterministic model structure leads to inconsistent estimates.

In duration models, the heterogeneity usually arises as a result of functional form misspecification. The misspecification is often due to an exclusion of a significant explanatory variable in the function form or lack of important unobservable variables. In linear regression, the omission of significant variables results in inconsistent estimates. Similarly, ignoring the heterogeneity in a duration model also leads to specification errors and inaccurate inferences of the covariate effects (Heckman & Singer, 1984; Lancaster, 1985).

A common method to account for heterogeneity is to include a disturbance term in the function form. The specification of this term is based on the distribution of the unobserved heterogeneity across individuals in the population. Similar to a normal disturbance term in linear regression, a random term following gamma distribution is usually included in the duration model to accommodate the heterogeneity. The gamma distribution is selected mainly because the property of this distribution provides a convenient close-form in computing log-likelihood function (Bhat, 1996). The analysis presented in this chapter adopts this approach to accommodate unobserved heterogeneity.

5.2 MODELING SURVEY PARTICIPATION DURATION WITH TRIP FREQUENCIES AS EXOGENOUS VARIABLES

5.2.1 Model with No Heterogeneity

In this model, the hazard function for a household *i* terminating its participation in the panel at time *t* is defined as:

$$l_i(t) = l_0(t)\exp(b' x_i), \qquad (5\text{-}8)$$

where x_i denotes a vector of covariates for individual *i*, and β denotes a corresponding vector of coefficients. The covariates that may affect the duration process include household demographics such as household income, characteristics of survey feature such as different sampling groups from which a household was selected, and/or survey performance indicator such as whether a survey unit has a variable missing. We first consider the case without time-

varying covariates. Let t_i represent the continuous time. The survival function then can be derived from equation (5-5),

$$S(t_i) = \exp\left(-\int_0^{t_i} l_i(t)dt \right) = \exp\left(-\exp(b'x_i)\int_0^{t_i} l_0(t)dt \right). \qquad (5\text{-}9)$$

It is assumed that the baseline hazard rate remains as a constant in each time interval. Thus, for any t_i in the time interval between wave k and wave $k+1$, the survival function is written as

$$S(t_i) = \exp\left[-\sum_{u=1}^{k} l_0(u)\exp(b'x_i) \right] = \exp\{-\exp[\ln\Lambda_0(k) + b'x_i]\}, \qquad (5\text{-}10)$$

where $\Lambda_0(k)$ is the integrated hazard which equals $\sum_{u=1}^{k} l_0(u)$.

Let a^k represent the time interval between wave k and wave $k+1$. It is assumed that the hazard rate remains a constant during each time interval. Let d_i denote the survey participation duration for household i. Using the definition of the survival function, the probability of household i not responding to the survey after it had stayed in the survey for k waves can be expressed as follows,

$$\begin{aligned}\Pr(d_i = k) &= \Pr(t_i \geq a^{k-1}) - \Pr(t_i \geq a^k) \\ &= \exp\{-\exp[\ln\Lambda_0(k-1) + b'x_i]\} - \exp\{-\exp[\ln\Lambda_0(k) + b'x_i]\} \\ &= G[\ln\Lambda_0(k-1) + b'x_i] - G[\ln\Lambda_0(k) + b'x_i],\end{aligned} \qquad (5\text{-}11)$$

where $G(z) = \exp(-\exp(z))$.

The probability takes the form of the ordered response model. The $\ln\Lambda_0(k)$ term in equation (5-11) acts as a threshold with an order of

$\ln \Lambda_0(k) \le \ln \Lambda_0(k+1)$ for all k's. For those households who returned the survey diaries in the last wave, we do not observe whether or not these households will continue responding to the next wave. Therefore, the probability for a censored household i is,

$$\Pr(d_i \ge k) = G[\ln \Lambda_0(k) + b' x_i]. \qquad (5\text{-}12)$$

This dissertation uses seven-wave PSTP data for empirical analysis so the maximum wave is seven and $K = 7$. Consequently, a total of eight corresponding integrated baseline hazard rates appear in the model, as demonstrated in Figure 5-1. Because the initial nonresponse households are not available for the analysis, all the households in the data set at least attended one wave of the survey. The data indicate that $\Pr(d_i \ge 1) = G[\ln \Lambda_0(0) + b' x_i] = 1$ for any household. Therefore, $\ln\Lambda_0(0)$ is set to be $-\infty$ to guarantee a probability of one. In the mean time, all of the households who attended the whole seven-wave survey are censored. No available information can determine how many households would participate for exact seven waves and how many others would go for more than seven. As a result, $\ln\Lambda_0(7)$ cannot be identified. Thus, six parameters are estimated for the non-parametric integrated baseline hazard rates.

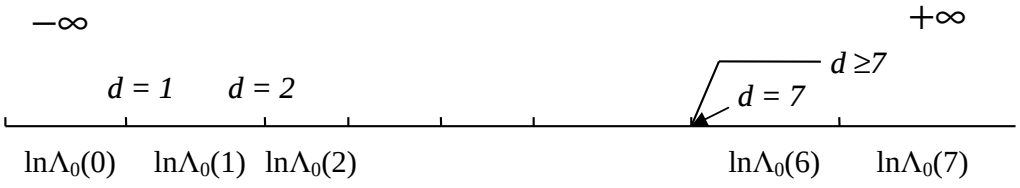

Figure 5-2: Layout of the non-parametric baseline hazard rates

The model is estimated using MLE. The log-likelihood function of this model takes a form that is similar to the ordered response model. It can be written as

$$\mathrm{LL} = \sum_{i=1}^{N}\sum_{k=1}^{K} M_{ik} \cdot \ln\left\{ G\left[(\ln \Lambda_0(k) - b' x_i)\right] - (1 - C_i) \cdot G\left[(\ln \Lambda_0(k+1) - b' x_i)\right]\right\} \tag{5-13}$$

where $M_{ik} = \begin{cases} 1 & \text{if household } i \text{ has participated for } k \text{ waves} \\ 0 & \text{otherwise} \end{cases}$,

and $C_i = \begin{cases} 1 & \text{if household } i \text{ is censored} \\ 0 & \text{otherwise} \end{cases}$.

Based on the definition of hazard rate, the estimated baseline hazard for a household staying in the panel for k waves is computed using the following formula,

$$l_0(k) = \frac{\Pr(t = k)}{\Pr(t \geq k)} = \frac{G[\ln \Lambda_0(k+1)] - G[\ln \Lambda_0(k)]}{G[\ln \Lambda_0(k)]}. \tag{5-14}$$

When including time-varying covariates in the model, the model formulation changes slightly. The change is mainly due to the integral with regard to time in the survival function. When the covariates vary over time, after

household *i* have participated in the survey for *u* waves the hazard function can be written as,

$$l_i(u) = l_0(u)\exp(b'x_{iu}), \tag{5-15}$$

where x_{iu} is a vector of covariates and β is a vector of parameters to be estimated and these parameters do not vary over time. Consequently, for any *t* in the time interval between wave *k* and wave *k+1*, the survival function is expressed as

$$S(t_i) = \exp\left[-\sum_{u=1}^{k}\left(l_0(u)\exp(b'x_{iu})\right)\right] = \exp\left[-\sum_{u=1}^{k}\exp(h_u + b'x_{iu})\right]. \tag{5-16}$$

where $h_u = \ln\left(l_0(u)\right)$. With the new formula of the survival function, the probability for a household *i* remaining in the survey for *k* waves is re-written as

$$\Pr(t_i = k) = \exp\left[-\sum_{u=1}^{k-1}\exp\left(h_u + b'x_{iu}\right)\right] - \exp\left[-\sum_{u=1}^{k}\exp\left(h_u + b'x_{iu}\right)\right]. \tag{5-17}$$

For a censored household, the probability is

$$\Pr(t_i \geq k) = \exp\left[-\sum_{u=1}^{k}\exp\left(h_u + b'x_{iu}\right)\right]. \tag{5-18}$$

The corresponding changes are made to the log-likelihood function to incorporate the time-varying covariate effects.

5.2.2 Model with Gamma Heterogeneity

In this model, a random term is introduced to represent the unobserved heterogeneity. Following equation (5-15), the hazard function is re-defined as

$$l_i(u) = l_0(u)\exp(b'x_{iu})w_i, \tag{5-19}$$

where w_i follows a gamma distribution with a mean of one and a variance of s^2 that need to be estimated. The conditional probability of attending k waves with no censoring is expressed as

$$\Pr(t_i = k \mid w_i) = \exp\left[-\sum_{u=1}^{k-1}\exp(h_u + b'x_{iu})w_i\right] - \exp\left[-\sum_{u=1}^{k}\exp(h_u + b'x_{iu})w_i\right]. \tag{5-20}$$

The unconditional probability is obtained by taking an integral of conditional probability over the distribution of w_i, then,

$$\Pr(t_i = k) = \int_0^{\infty}\left\{\exp\left[-\sum_{u=1}^{k-1}\exp(h_u + b'x_{iu})w_i\right] - \exp\left[-\sum_{u=1}^{k}\exp(h_u + b'x_{iu})w_i\right]\right\} f(w_i)\,dw_i \tag{5-21}$$

where $f(w_i)$ is the probability density function of w_i. Johnson and Kotz (1970; also see Bhat and Steed, 2001) showed that, by using the moment-generating method, the following equation can be derived from equation (5-21)

$$\Pr(t_i = k) = \left\{1 + s^2\left[\sum_{u=1}^{k-1}\exp(h_u + b'x_{iu})\right]\right\}^{-s^{-2}} - \left\{1 + s^2\left[\sum_{u=1}^{k}\exp(h_u + b'x_{iu})\right]\right\}^{-s^{-2}}. \tag{5-22}$$

For the censored cases, the probability for household *i* attending at least *k* waves of the survey is expressed as

$$\Pr(t_i \geq k) = \left\{1 + s^{2}\left[\sum_{u=1}^{k} \exp(h_u + b' x_{iu})\right]\right\}^{-s^{-2}}. \qquad (5\text{-}23)$$

Similarly, the log-likelihood function for the parameter estimation can be written as

$$\mathrm{LL} = \sum\sum\left(M_{ik} \ln\left\{\left[1 + s^{2}\left(\sum_{u=1}^{k-1} \exp(h_u + b' x_{iu})\right)\right]^{-s^{-2}} - (1 - C_i)\left[1 + s^{2}\left(\sum_{u=1}^{k} \exp(h_u + b' x_{iu})\right)\right]^{-s^{-2}}\right\}\right). \qquad (5\text{-}24)$$

where M_{ik} and C_i follow the definitions in equation (5-13). The models are estimated using the GAUSS econometric package for its conveniently implemented MAXLIK (maximum likelihood estimation) module. The analytical gradient function is also coded to achieve a faster convergence.

5.3 DATA SETS FOR MODEL ESTIMATION AND VALIDATION

The data used for this study consists of a total of 4802 households. These households are randomly divided into two data sets: calibration set and validation set. The calibration data is used for model estimation and the validation data is used for model evaluation. There are a total of 3348 households in the calibration data, about 70% of the total observations. The average participation duration for households is 2.82 waves in the calibration data and in the validation data the average is 2.86 waves. The descriptive statistics of the survey participation

duration are presented in Table 5-1. The sample fractions show no significant different in these two data sets. We developed a set of proportional duration models and a binary logit model using the calibration data and the comparison are based on the validation data.

Table 5-1: Households' survey participation duration in calibration and validation sets

	Duration	**Calibration set**	**Validation set**
Number of households with no censoring (sample fraction)	One wave	612 (18.3%)	257 (17.6%)
	Two waves	615 (18.4%)	284 (19.5%)
	Three waves	274 (8.2%)	97 (6.7%)
	Four waves	160 (4.8%)	73 (5.0%)
	Five waves	207 (6.2%)	89 (6.1%)
	Six waves	94 (2.8%)	33 (2.3%)
Number of households with censoring (sample fraction)	One wave	469 (14.0%)	204 (14.0%)
	Two waves	318 (9.5%)	125 (8.6%)
	Three waves	84 (2.5%)	40 (2.8%)
	Four waves	84 (2.5%)	51 (3.5%)
	Five waves	64 (1.9%)	31 (2.1%)
	Six waves	48 (1.4%)	21 (1.4%)
	Seven waves	319 (9.5%)	149 (10.2%)
Total		3348 (100%)	1454 (100%)

5.4 EMPIRICAL RESULTS

Summary statistics for the hazard models are shown in Table 5-2. A total of 26 parameters, including 6 baseline parameters and 20 covariate coefficients, are estimated in the model with no heterogeneity. The model with Gamma heterogeneity consists of 28 estimated parameters. The log-likelihood value at convergence for the model with no heterogeneity is –3665.190 and –3649.688 for the model with Gamma heterogeneity. The likelihood ratio test shows that the restricted model (the one with no heterogeneity) is rejected.

Besides the log-likelihood value at convergence, Table 5-2 also presents another two log-likelihood values as benchmarks. One is the log-likelihood value with baseline parameters only and the other is the log-likelihood value at zero. The log-likelihood value with baseline parameters only refers to the case in which no covariate effects and heterogeneity are accommodated. The only sample information used in this case is the temporal dynamics which are captured by the baseline hazard parameters. The log-likelihood value at zero corresponds to the case when no sample information is used for model estimation. It is assumed that households would terminate their survey participation in each time interval with equal probability. The adjusted likelihood ratio index is then defined as

$$\bar{r}^2 = 1 - \frac{\text{log - likelihood at convergence} - \text{total number of parameters}}{\text{log - likelihood at zero}} \tag{5-25}$$

The adjusted likelihood ratio index shows that the model with Gamma heterogeneity fits better with the calibration data than the one with no heterogeneity incorporated, but the improvement is marginal (0.436 vs. 0.434).

Table 5-2: Summary statistics for the hazard models

Summary statistic	**Semi-parametric proportional models**	
	No heterogeneity	**Gamma heterogeneity**
Number of observations	3348	3348
Number of baseline parameters	6	6
Number of unobserved heterogeneity parameters	0	1
Number of covariates	20	21
Total number of estimated parameters	26	28
Log-likelihood at convergence	-3665.190	-3649.688
Log-likelihood with baseline parameters only	-4416.581	-4416.581
Log-likelihood at zero	-6514.907	-6514.907
Adjusted likelihood ratio index	0.434	0.436

5.3.1 Covariate Effects

Earlier studies (Meurs and Ridder, 1997) found that household demographics have interpretation power on the survey participation decision. Meanwhile, from a behavioral point of view, the combination of survey burden, time constraint, and the households' commitment to the survey subject may explain why some households respond to the travel activity survey and some others do not. To accommodate various aspects of the decision making process,

the independent variables initially considered in our model specification fall into three categories. These three categories are: household demographics, survey attributes, and trip making characteristics.

The household demographic variables include household size, household life cycle type, household income, age structure, and household vehicle ownership. Other variables, such as the number of eligible household members to fill out the travel diary, are incorporated to measure the survey burden. The survey-related attributes include which sampling group households belonged to, whether or not there was an attitude survey accompanied with the activity survey, and whether or not there is an imputed item nonresponse. In addition, the models examine the impact of trip frequencies on the household travel survey participation by including the number of home-based work, home-based non-work and non-home based trips as exogenous variables. Table 5-3 shows the estimated coefficients of household demographic variables and the parameters for other variables are presented in Table 5-4 for models with no heterogeneity and Gamma heterogeneity.

Comparing the two models, the coefficients have the same sign in both models. Variables with a positive sign indicate that it would be more likely for a household to terminate the survey participation as these variables increase. Similarly, a negative sign of a coefficient demonstrates that the household would be more likely to stay in the survey as the variable goes up. Quantitatively, Heckman and Singer (1984) pointed out that failure to incorporate the heterogeneity would lead to a bias toward zero for the effect of external

covariates. Our empirical results show that the restricted model does experience this type of bias toward zero. The coefficients estimated in the restricted model are closer to zero than those in the model with Gamma heterogeneity. The estimate for the variance parameter s^2 is 0.5256. This parameter is significantly different from zero with a t-value of 4.1476, rejecting the null hypothesis of no unobserved heterogeneity across households.

Table 5-3: Covariates effect of household demographics

Independent Variable	No Heterogeneity			Gamma Heterogeneity		
	Coefficient	t-Value	Significant Level	Coefficient	t-Value	Significant Level
Households with all children between the age of 6 to 17	-0.195	-2.434	0.0149	-0.2514	-2.523	0.0116
Households with no children and one adult under the age of 35	0.3874	3.239	0.0012	0.4825	3.076	0.0021
Households with no children and one adult between the age of 36 to 64	-0.5338	-4.922	0	-0.5335	-3.927	0.0001
Households with no children and one adult older than 65	-0.865	-6.641	0	-0.8792	-5.393	0
Households with no children and two+ adults under the age of 35	0.3162	3.092	0.002	0.4731	3.557	0.0004
Households with no children and two+ adults between the age of 36 to 64	-0.343	-4.406	0	-0.3786	-3.881	0.0001
Households with no children and two+ adults older than 65	-1.3193	-12.636	0	-1.5751	-11.4	0
Number of workers	-0.9727	-26.63	0	-1.1416	-20.58	0
Household income: 25K to 45K	0.1945	3.914	0.0001	0.234	3.852	0.0001
Split households	-	-	-	0.4736	1.928	0.0539

Table 5-4: Covariate effect of survey burden, sampling group, and others

Independent Variable	No Heterogeneity			Gamma Heterogeneity		
	Coefficient	**t-Value**	**Significant Level**	**Coefficient**	**t-Value**	**Significant Level**
Households entering the panel in wave 2	0.5916	7.561	0	0.8113	6.649	0
Households entering the panel in wave 3	0.4475	5.176	0	0.512	4.675	0
Households entering the panel in wave 4	0.778	10.043	0	1.0487	9.014	0
Households entering the panel in wave 5	0.9785	9.578	0	1.2033	8.719	0
Households entering the panel in wave 6	1.2562	12.016	0	1.4742	11.278	0
Number household members who filled out the travel diary	0.3458	8.174	0	0.5367	7.835	0
Carpooling sample group	0.1218	1.653	0.0984	0.1462	1.507	0.1318
Household demographic attributes (other than income) are imputed	0.2614	5.017	0	0.2824	4.23	0
Household income is imputed	-0.2968	-4.335	0	-0.4589	-4.821	0
Number of home-based work trips	0.0383	4.033	0.0001	0.0343	2.86	0.0042
Number of home-based non-work trips	-0.0318	-6.106	0	-0.0398	-6.246	0
σ^2	-	-	-	0.5256	4.1476	0

5.3.1.1 Impact of Household Demographic Variables

The modeling results show that the household life cycle has a significant influence on the survey participation decision. Seven dummy variables representing different stages of household life cycle are included in the models and all of them are significant. For identification reason, the dummy variable for households with any children under the age of 5 is not included in the models. It is used as benchmark to evaluate the impact of household life cycle.

The model results show that households with all children in age 6 to 17 are more likely to continue their survey participation comparing to households with children under age 5. For no-children households, it appears that households with young adults are more likely to drop out of the panel survey and households with older adults tend to continue their participation. The households in the under 35 age group are the most likely to terminate their survey participation, while the households with adults older than 65 are the most likely to continue responding to the survey. The differences among households in various age groups may be a combined result of social concern, time constraint, and life stability. The younger households can often be categorized as a group more active and with more life pressure. On the contrary, the older or retired households are more stable, with less time constraint, and are probably more concerned with public issues. Consequently, it is more likely for them to constantly participate in the survey than the younger households.

We initially included the number of household members by age in the models, but these variables turned out to be statistically insignificant. The

insignificance is probably because the age information is well incorporated in the household life cycle variables.

The number of workers in the household also significantly affect the survey participation duration. It appears that the more the number of household members who are employed, the more likely it is that the household will continue to participate in the survey. This result does not quite fit our initial expectation. We expected that employed household members might have more time constraints and therefore, it might be more difficult for them to complete the survey. However, since most of the workers commute during peak hours, they may have more personal experiences with urban congestion, which often occurs during peak hours. These experiences may raise their concern about traffic problems and their concern may make their households more willing to participate in the panel survey. Furthermore, workers generally have a more regular activity schedule during weekdays than unemployed people. This regularity may make it easier for them to record the activities in the travel diary.

The regularity in workers' daily activity pattern may depend on employment type and household/personal demographics. It is beyond the scope of this dissertation to analyze the regularity of workers' activity patterns. However, it should be noted that, quantitatively, the absolute value of the coefficient for the number of workers (0.9727 in the model with no heterogeneity and 1.1416 in the model with Gamma heterogeneity) is larger than any other non-dummy variables. In addition, the employed household member is often the head of the household and he/she may ultimately determine whether or not to

participate in the panel survey. Thus, further examination on the impact of household workers' employment type and activity pattern may be worthwhile to better understand households' decision on survey participation.

The initial model specification includes three dummy variables representing low-median, high-median, and high income groups. Household income ranges from $25k to $ 45k for the low-median income group, $45k to $75k for the high-median income group, and above $75k for the high income group. The model result shows that low-median income has a positive impact on the hazard rate, indicating that households with income ranging from $25k to $45k are more likely to decline the survey request. Meanwhile, we found that high-median and high income have an opposite impact on the hazard rate. However, the estimated parameters for these two income groups are not statistically significant, so they are removed from final model specification.

Throughout the survey period, some households had one or two split households. A common reason for the split households is that a young adult moved out to college or got married. The model with Gamma heterogeneity suggests that split households are less willing to continue participating in the survey. It is probably because the initial decision for the survey participation was made at their parents' house and they are less committed to the survey.

In transportation literature vehicle ownership is an important factor associated with trip making behavior. The vehicle ownership is often viewed as a mobility indicator. Therefore, it might affect households' decision to participate in travel surveys. Our model results show that the number of vehicles owned by

households does not have significant impact on the survey participation duration. The insignificant impact is probably because of the collinearity between vehicle ownership and other mobility indicators such as trip frequency which are also included in the model. In addition, the sample households were recruited from different user groups (regular SOV users, carpoolers, and transit users). The choice-based sample groups also reflect household vehicle ownership to some extend. The impact of these variables is presented in the next section.

5.3.1.2 Impact of Survey Attributes and Trip Characteristics

In the sample data households started their travel survey participation at different points in time. The majority of the households (36%) participated in the survey since wave 1 as a result of random phone calls and volunteers in transit-user group. Other households entered the survey as refreshment sample in the following waves (wave 2 to wave 7). These households were recruited by carefully matching household life cycle and choice-based sample group with those who declined the survey request. One hypothesis is that these refreshment households are intrinsically more vulnerable than those recruited in wave 1 in terms of continuous survey participation. In addition, when they started their survey participation in different waves, their first experience about survey differed from each other (for example, some started with travel attitude survey and some did not). The hypothesis is that households' first-time experience in the survey significantly affects their participation decision in the later waves. We include five dummy variables representing the different start points of

households' survey participation in the model to test these hypotheses. The model results do show that households recruited in the later waves are more likely to terminate their survey participation comparing to those recruited in wave 1, especially for households recruited in wave 6. The higher hazard rate for households who started their survey participation in wave 6 is probably because of two reasons. First, it may be due to the way in which refreshment households were recruited for wave 6. Unfortunately, this information is not available for further examination. Second, differences in survey operation may contribute to the higher likelihood of terminating survey participation. For instance, unlike the other waves in which data were collected during the fall, wave 6 was conducted during the summer time. Different data collection seasons may lead to the model results.

The activity survey requires every household member older than 15 years to fill out the travel diary. The results show that the more household members that are eligible to fill out the diary, the less likely the household will participate in the survey for more waves. The lower likelihood must be due to the extra cooperation needed among household members. It illustrates different aspects of the decision process within households. On one hand, the decision may heavily depend on one particular household member. It might be the chief member, or the one who received the call in the initial random phone contact. On the other hand, the participation requires all the members' commitment. More eligible members required to fill out the diary generally implies more difficulties to reach an agreement about survey participation within a household.

Throughout the PSTP survey, some waves collected not only travel diaries, but also a travel attitude questionnaire. The travel attitude survey requires the households' extra efforts to finish, so more households may stop responding to the survey due to extra survey burden. Our models show that after a wave accompanied with the attitude survey, the households are more likely not to respond in the following wave. However, the impact of accompanied attitude survey is not statistically significant so we did not include this variable in the final model specification.

The sample units in the PSTP are choice-based samples belonging to each of the three sampling groups: regular SOV user, carpooler, and transit user. The households categorized in the transit-user group, instead of being randomly selected, were selected from those who indicated in previous transit surveys that they are willing to participate in the future survey. However, there is no evidence in the models indicating that households in transit-user sample group participated for more waves than regular driving-alone households. Maybe the survey which they meant to respond to is a transit survey instead of a general purpose household travel survey since transit users are generally more interested in transit surveys that may help improve transit service. Furthermore, they may be less interested in completing a household survey which requires other household members' participation. It is also possible that their initial voluntary intention of survey participation was negated by the time constraints and inconveniences imposed on transit users.

Instead of the transit-user group, the model results show that the regular carpooler group has a positive impact on the hazard rate, which indicates that the carpooling households are more likely to terminate the survey participation compared to regular SOV and transit-user households. It is difficult to explain why the regular carpooling households have a higher hazard rate. It might be because of the necessary coordination among carpoolers which imposes more time constraints on them.

Previous studies and intuition both suggest that item nonresponse may indicate a potential unit nonresponse in the future. Our model results confirm this implication. Two item nonresponse indicators are significant in the final models. One is a dummy variable for the existence of a missing value for income; the other is an indicator representing other household demographic variables with a missing value. As expected, a missing household demographic variable (other than income) leads to a higher hazard rate. A partially completed survey questionnaire usually reflects the household's reluctance to respond and it often occurs that the household completely drop out of the survey for the following wave.

Another interesting finding is that a missing income variable does not imply the same pattern as other missing variables. In fact a missing income variable would lead to a lower hazard rate, suggesting that these household are more likely to remain in the survey. The different covariate effects of missing value indicators may be due to two reasons. First, a missing income variable generally takes place independently of other missing values, while the rest of the

demographic variables often have missing values at the same time. Second, many households are reluctant to provide information on income for different reasons. There are many more households with income missing than any other demographic variables. Missing income in the survey is so common that it is not obvious what the phenomenon itself suggests for the future survey participation.

Similarly, opposite effects are found for the number of home-based work trips and non-work trips, although the absolute values of the coefficients are small (0.0343 and -0.0398 in the model with Gamma heterogeneity). The results show that the households with more home-based work trips made during the survey period tend to decline the survey request. On the contrary, each extra home-based non-work trip will slightly lower the hazard rate and the households are likely to continue their participation for the next wave. The number of non-home-based trips is not significant in the final model specifications.

If trip rate is considered as a survey burden indicator, it is expected that more trips will lead to a higher likelihood for a household terminating the survey participation. Our results on home-based work trips do support this hypothesis. It is the negative impact of home-based non-work trips that deserves extra attention. A possible explanation is that a household often makes more home-based non-work trips when the household members have more time. In other words, more home-based non-work trips imply a less restricted time constraint imposed on the household. In addition, the time flexibility overcomes the hassle of reporting every travel activity so that the more non-work trips actually result in a higher

probability of responding to the survey. We will further examine the relationships among survey participation, survey burden, and trip frequency in Chapter 6.

5.3.2 Baseline Hazard Rate

Figure 5-2 shows the baseline hazard rate obtained from the model with no heterogeneity and Figure 5-3 demonstrates the baseline hazard rate for the model with Gamma heterogeneity. The baseline hazard rates exhibit the same pattern in both diagrams, although the hazard rates obtained from the model with Gamma heterogeneity are higher than those of the model with no heterogeneity.

The diagrams show that there is positive duration dependence in the hazard function, i.e., the more waves a household has participated in, the more likely it would stop responding to the next wave. However, the baseline hazard rate does not increase monotonically. It is indicated that households' participation in the panel surveys can be divided into three stages. The break points are after households have participated for two waves and five waves. The first rising in the baseline hazard rate happens after households have been with the survey for two waves. Afterward the baseline hazard remains at a steady level (in the model with no heterogeneity, it decreases a little) until households approach their fifth survey participation. It is the most stable stage in terms of duration dynamics. After a household has participated for five waves, the baseline hazard rate abruptly increases to nearly one, indicating that it is ***almost sure*** that these households will not respond to the sixth wave if no covariates effect is considered. The abrupt increase in the hazard rate may suggest a panel fatigue point.

The dynamics in the baseline hazard rate suggest that more survey operation and administration efforts be carried out when households have been participating in the survey for two waves and five waves because of the substantial increase in the probability of terminating the survey participation. Often faced with budget constraints in practice, survey operators may choose alternatives of reducing sample size and rotating sample units to achieve a higher response rate after households reach a panel fatigue point.

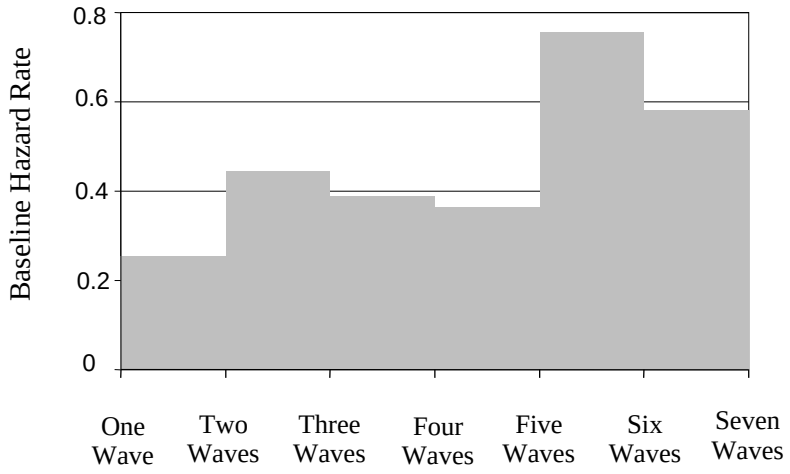

Figure 5-2: Baseline hazard rate (no heterogeneity)

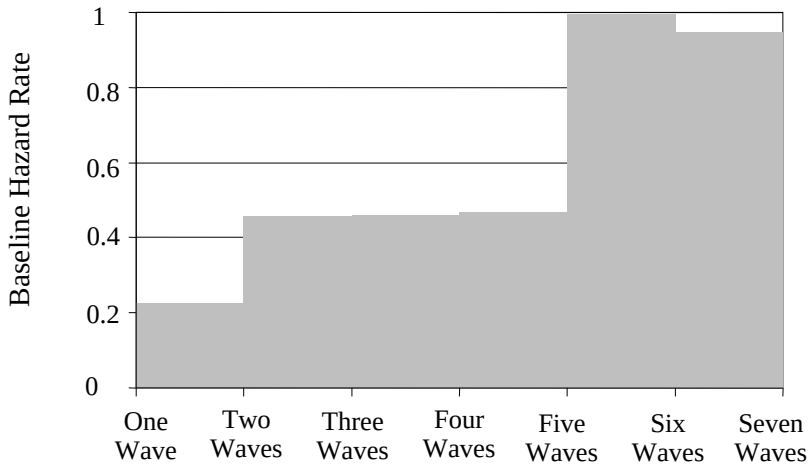

Figure 5-3: Baseline hazard rate (Gamma heterogeneity)

5.5 COMPARISON OF HAZARD-BASED DURATION MODEL WITH DISCRETE CHOICE MODEL

Since discrete choice model is commonly used for survey nonresponse analysis, we estimate a binary logit model and compare it with the proportional hazard duration model (with no heterogeneity) presented in Section 5.4. The comparison focuses on the parameter estimates and the fitness of the model.

In the binary logit model, the dependent variable is whether a household will stop responding to the survey in the next wave. It is equal to 1 if a household continues to participate in the survey and 0 otherwise. The estimated coefficients for the logit model are presented in Table 5-5. The log-likelihood value at convergence is -3533.533 and the log-likelihood at zero is -5580.528. The adjusted likelihood ratio index is 0.3629 which is computed following equation (5-24).

Table 5-4 compares the sign of estimated parameters in both models. In the hazard duration model, a positive coefficient suggests a higher likelihood of terminating the survey participation with an increase in the independent variable. A positive coefficient in the logit model reflects an opposite effect. It indicates that the household is more likely to respond to the survey for the next wave when the variable increases. If the findings of these two models are consistent with each other, the corresponding parameters should have opposite signs. Comparing the sign of parameters in the models, we found that the model estimates are consistent for most of the parameters. There are a few variables that are found significant in one model but insignificant in the other. For instance, the model

results show that households in carpooling sample group are more likely to stop responding to the survey in the duration model, but in the logit model the impact is not statistically significant. Similar phenomenon occurs to households in different income groups.

The only variable whose impact on survey response is found inconsistent in two models is the dummy variable for households with a missing income variable. The duration model finds that households with a missing value of income actually stay longer with the survey, while the logit model shows that a missing income variable reflects households' unwillingness to participate in the following wave, although the magnitude of the estimated coefficient is relatively small (-0.4101). We should point out that even though the duration model and the logit model found different effects of missing values in income variable, both model results support the hypothesis that the impact of a missing value in income is significantly different from that of a missing value in other household demographic variables.

Table 5-5: Estimated coefficients for the logit model

Independent Variable	Coefficient	t-value	Significant level
Constant	1.7404	9.428	0
Dummy variable for wave 2	-0.9054	-7.138	0
Dummy variable for wave 3	-0.4149	-3.024	0.0025
Dummy variable for wave 4	-0.8255	-6.534	0
Dummy variable for wave 5	-1.5486	-12.636	0
Dummy variable for wave 6	-1.5693	-12.828	0
Household type			
Households with all children between the age of 6 to 17	0.1926	1.843	0.0654
Households with no children and one adult under the age of 35	-0.6538	-3.918	0.0001
Households with no children and one adult between the age of 36 to 64	0.6408	4.474	0
Households with no children and one adult older than 65	1.0175	6.204	0
Households with no children and two+ adults under the age of 35	-0.5763	-4.280	0
Households with no children and two+ adults between the age of 36 to 64	0.5766	5.854	0
Households with no children and two+ adults older than 65	1.7540	13.645	0
Other household demographics			
Number of workers	0.9819	20.166	0
Number of adults	-0.3784	-2.982	0.0029
Household income: 45K to 75K	0.1225	1.702	0.0888
Household income: >75K	0.2039	1.937	0.0527
Survey attributes and trip frequency			
Number household members who filled out the travel diary	-0.2222	-2.067	0.0388
Household demographic attributes (other than income) are imputed	-5.0996	-7.325	0
Household income is imputed	-0.4101	-4.019	0.0001
Number of home-based work trips	-0.0322	-2.423	0.0154
Number of home-based non-work trips	0.0377	5.593	0
Log-likelihood at convergence	-3533.53		
Log-likelihood at zero	-5580.528		
Adjusted likelihood ratio index	0.3629		

Table 5-6: Comparison of the sign of estimated parameters

Independent Variable	Model	
	Hazard duration	Logit
Household type		
Households with all children between the age of 6 to 17	Negative	Positive
Households with no children and one adult under the age of 35	Positive	Negative
Households with no children and one adult between the age of 36 to 64	Negative	Positive
Households with no children and one adult older than 65	Negative	Positive
Households with no children and two+ adults under the age of 35	Positive	Negative
Households with no children and two+ adults between the age of 36 to 64	Negative	Positive
Households with no children and two+ adults older than 65	Negative	Positive
Other household demographics		
Number of workers	Negative	Positive
Number of adults	Insignificant	Negative
Household income: 25K to 45K	Positive	Insignificant
Household income: 45K to 75K	Insignificant	Positive
Household income: >75K	Insignificant	Positive
Survey attributes and trip frequency		
Number household members who filled out the travel diary	Positive	Negative
Carpooling sample group	Positive	Insignificant
Household demographic attributes (other than income) are imputed	Positive	Negative
Household income is imputed	Negative	Negative
Number of home-based work trips	Positive	Negative
Number of home-based non-work trips	Negative	Positive

The performance of these two models is evaluated using different measures. One of the goodness-of-fit criteria is the adjusted likelihood ratio index ($\bar{r}^2$). The adjusted r^2 can be viewed in a way as adjusted R^2 in regression analysis. The adjusted likelihood ratio index is 0.434 for the duration model and 0.363 for the logit model for calibration data, suggesting the hazard duration model fits the sample data better than the logit model. In addition, we use dissagregate measures to evaluate the models on validation data to verify that the models do not over-fit the sample data. The dissagregate measures of two models are presented in Table 5-7. The predictive log-likelihood value is computed by applying the estimated parameters to the validation data, and so as predictive adjusted log-likelihood ratio index. For both models, the predictive adjusted log-likelihood ratio index values are somewhat lower than those in calibration data (0.3834 vs. 0.434 and 0.268 vs. 0.363). It shows that the duration model is more stable and outperforms the logit model.

Table 5-7: Goodness-of-fit measures in validation data

	Hazard duration model (no heterogeneity)	**Binary logit model**
Log-likelihood at zero	-2829.353	-2448.195
Log-likelihood at sample shares	N/A	-2005.388
Log-likelihood at baseline hazard rates only	-1898.26	N/A
Predictive log-likelihood	-1746.92	-1770.28
Number of observations	1454	3532
Total number of parameters	26	22
Predictive adjusted log-likelihood ratio index	0.3834	0.268

5.5 SUMMARY

This chapter formulates a proportional hazard duration model to analyze survey participation duration in household panel surveys. The model structure has advantages in accommodating duration dependence, time-varying covariates effect, as well as the censored cases. Meanwhile, a disturbance term is added to the model to account for heterogeneity across households.

The empirical results indicate that positive duration dependence is observed in the panel surveys. However, the baseline hazard rate is not monotonically increasing across waves. There are two abrupt upsurges in the baseline hazard rate after households have been participating for two waves and five waves respectively. After these two rapid increases the baseline hazard rate remains at a steady level or slightly decreases. It should be noted that the baseline hazard rate for a household after responding to the survey for five waves is not significantly different from 1, suggesting households reach a panel fatigue point.

The coefficients of covariates effect suggest that household demographics and survey features have significant impacts on survey participation. It is found that households with older adults and more workers are more likely to continue participating in the panel survey. It is also revealed that more survey burden generally leads to a lower response rate. However, when survey burden is interpreted as different indicators, the results are not always consistent. For instance, more household members required to fill out travel diaries would lead to a higher likelihood for the household to terminate the survey participation, while for another survey burden indicator, the number of home-based non-work trips,

the model results show that households making more home-based non-work trips actually stay longer with the survey.

We also compared the duration model results with a binary logit model. The comparison shows that most of the parameter estimates reveal a similar impact on survey participation in both models. Furthermore, we use different criteria to evaluate the performance of the models. The model evaluation shows that the duration model fits the data better and more stable.

The next chapter will discuss more on survey burden measurement and investigate the correlation between survey participation duration and trip frequency.

Chapter 6 Capturing Observed and Unobserved Factors Associated with Survey Participation and Trip Frequency

In this chapter, a multi-level modeling system is proposed to investigate survey participation duration, trip frequency, as well as the correlation between the two in longitudinal household travel survey. The models presented in Chapter 5 suggest that the number of trips a household made during the survey period, often viewed as an indicator for survey burden, explicitly affect the household's survey participation duration. The results lead us to test a more realistic behavioral hypothesis that the correlation between survey participation duration and trip frequency is not only explicit but also endogenous. Ignorance of this endogenous correlation will result in biased parameter estimates for both the duration model and trip frequency model. This chapter is organized as follows. Section 6.1 describes various survey burden measures since trip frequency is considered as one of them for travel surveys. Section 6.2 discusses the modeling considerations, including model structure for panel data analysis (for trip frequency model) and the implementation of observed and unobserved heterogeneity. The formulation of the model system is presented in Section 6.3. Section 6.4 describes the simulated estimation method. Section 6.5 presents the model results, followed by a summary of the chapter.

6.1 MEASUREMENTS OF SURVEY BURDEN

It has long been believed that survey burden has negative impact on survey response, although previous studies do not always suggest so (Bogen, 1996; McCarthy and Beckler, 1999; Yu and Cooper, 1983; Heberlein and Baumgartner, 1978). A first step toward impact analysis of survey burden is to set up a quantitative measurement for it. The survey burden can be represented by the number of survey questions, the time to complete the questionnaire, or the number of contacts made to the survey participants.

In the case of panel data collection, the duration of the survey, i.e., the number of waves for which the survey units are expected to participate, certainly is a critical survey burden indicator for the survey units to determine whether or not to respond to the survey. However, in practice, this piece of information is not always clear even to the survey administrator due to many reasons, such as funding availability or shifts in public interest. For instance, when we obtained the seven waves of PSTP data from the Puget Sound Regional Council in October 2000, it was indicated that the last wave would be conducted in the fall of 2000. However, later at the Transportation Research Board annual meeting in January 2002, we were notified that an extra wave will be collected in the coming fall. It is not difficult to imagine that the sample units have no idea about how many waves a panel survey will last while they were initially contacted. Consequently, there is no way to examine how panel length is perceived by the households for their initial survey participation decision.

On the other hand, the information on the number of repetitive contacts may be more significant to households' initial decision than to the follow-up participation. After survey units have gone through the survey questionnaire once, their past experiences on the survey length may dominate the concern of repetitive contacts. This dissertation models the survey nonresponse conditional on household's initial response to the survey. Thus, the impact analysis of survey burden focuses on the measurement of survey length.

In general, travel surveys are conducted at household level. Different from some other household surveys in which one household member may be able to provide all the information, such as socio-economic characteristics, about other household members, household travel survey requires every eligible member's involvement to fill out the travel diary. The coordination effort within the household is one type of survey burden measurement. In the Puget Sound Transportation Survey, for example, every household member who is older than 15 needs to record his/her travel activities during the two-day survey period. The participation is a joint decision of the household and survey burden can be characterized by the number of eligible household members.

As the travel diary records eligible household members' travel activities, another indicator of survey burden is the number of travel activities that needs to be recorded in the travel diary. When a household member makes stops at different locations during the day, it is possible that he/she may quit reporting these activities due to the inconvenience or the trouble in recalling these activities. An essential way to reduce this type of survey burden is to adopt intelligent

survey devices such as Global Positioning System. With pre-programmed activity categories integrated in the device, survey participants just need to do the categorical selection while they are making each stop. Therefore, it would be much easier for the activity reporting process and survey burdens may not be a significant concern when households are making the participation decision. In the conventional paper-pen activity survey, common sense suggests that efforts to report each travel activity can be substantial and significant to the survey participation decision.

For transportation study, a general purpose of modeling survey nonresponse is to derive a weighting system in order to obtain consistent parameter estimates that capture various aspects of travel behavior. Therefore, the consistence of the attrition model is important not only to understand the cause of nonresponse but also to properly weight the observed survey records. Meanwhile, trip frequency, as an indicator of survey burden in travel activity surveys and one essential character of travel behavior, is of interest in many transportation studies and often is why a travel survey is initially conducted. If there is an endogenous correlation between nonresponse behavior and the subject of interest (trip frequency), the sequential estimation procedure, i.e., first modeling survey nonresponse and then trip frequency, often leads to biased estimates.

In this chapter, we formulate a multi-level model system that considers trip frequency and panel attrition simultaneously to further investigate the relationship among survey burden, survey participation and travel behavior. First,

we aim to examine the impact of survey burden on households' decision to participate in the survey. We use various measurements to represent survey burden. Second, we seek to identify key characteristics of survey non-respondents. A realistic and intuitive model structure is an essential to effectively segment survey non-respondents. More importantly, the results can be applied to the following waves in order to increase the response rate. With the same survey subject and data collection method, the attrition model is transferable across panel waves. Third, by modeling trip frequencies and panel attrition simultaneously, we intend to capture the internal correlation between them and characterize trip-making behavior across waves.

When studying the impact of survey burden on survey participation, from behavioral point of view, the analysis cannot be complete without considering the time constraint for different survey participants. For instance, employed household members need to go to work or a mother needs to drop her kid at school at certain time. These events are time-sensitive and therefore the survey participants may be reluctant to record these events in the travel diary due to the time constraint. It is always challenging to implement temporal or spatial constraints in conventional econometric models. A practical approach is to include explanatory variables, such as employment status or accessibility index, in the model specification to reflect the time or spatial constraints. An earlier modeling effort in this dissertation considered trip frequency as an exogenous variable in the survey duration model. It is interesting to note from the analysis that the rates of home-based work trips and home-based non-work trips have

totally opposite impacts on households' continuing participation in the survey. The different results for work and non-work trips may be due to the endogenous correlation among trip making and survey participation behavior. To further examine the relationship among survey participation and trip frequencies, we include two trip frequency equations for home-based work and home-based non-work trips respectively in the model system.

6.2 MODELING CONSIDERATIONS

6.2.1 Correlations among Trip Frequency and Survey Participation Duration

One key consideration for the multi-level model system is the correlation among the survey participation and travel activities. The correlation may be due to some observed factors, such as household income and other household demographic characteristics. The observed effect can be captured by introducing these socio-economic variables in the systematic component (i.e., V_{iq} in equation 3-1) of latent propensity function (or utility function). The number of trips is also included in the survey participation model as an independent variable. In addition, it is likely that the correlation may exist as a result of unobserved heterogeneity. A general approach is introducing a common disturbance term to reflect the mutual variation in the survey participation and trip frequency. The approach is equivalent to defining a variance-covariance matrix for the disturbance terms. The model formula implementing both observed and unobserved heterogeneity will be presented in Section 6.3.

6.2.2 Dynamic vs. Static Model for Panel Data

When modeling households' decision on continuing participation in the panel surveys as a duration process, we implement the time-varying covariates effect in the hazard function because it is observed that the household demographics change during the survey period. Simply implementing time-invariant effects in the model would be against intuition and would result in a loss of information. Similarly, households make different numbers of travel activities from wave to wave. Trip frequency, as an indicator of survey burden, cannot be identical across waves. Consequently, some modeling issues need to be considered when trip frequencies across waves are of interest in the model system.

The first consideration for the trip frequency model structure is dynamic vs. static model. The dynamic model can be in the form of a first order Markov chain model in which the trip rates in the following wave partially depend on the trip rates in the previous waves, while in a static model the trip making behavior is fully explained by the current household demographics. We select the static model structure in this study for the following reasons. First, previous studies examining the trip rates in panel waves has found that the household's trip making behavior is relatively stable from wave to wave (Kitamura and Bovy, 1987). The trip making behavior usually can be well explained by the household demographics. Second, the Puget Sound Transportation Survey is conducted about one or two years apart for each wave and the travel diary records household members' two-weekday travel activities. It is difficult to accommodate the day-

to-day variation using two-day data. Furthermore, travel activities are often influenced by season and unrecorded weather condition. Because of the day-to-day variation, advantages of the dynamic model may not be significant in this case. Third, if adopting the dynamic model, say, first-order Markov chain model, the estimation of the initial condition would tremendously complicate the model structure and estimation. Thus, the static model structure is utilized to model trip frequencies in this study.

In the model formulation, we adopt an ordered response choice model structure for trip frequencies. Since the formulation is applied to multi-wave panel data, disturbance terms are introduced to the trip frequency equation to accommodate the unobserved individual effect and the unobserved time effect. The hazard-based duration model is utilized to model the attrition behavior. The two models are connected with each other by the number of trips made by households during the survey period and common disturbance terms which accommodate for the unobserved heterogeneity.

6.3 MODEL STRUCTURE

The multi-level model system has three equations. One is the proportional hazard-based duration model with random coefficients which describes households' decision on continuing participation in the panel survey. The other two equations follow the random coefficient ordered probit choice model structure for home-based work and home-based non-work trips respectively.

By adopting the non-parametric baseline hazard function, the hazard-based duration model turns into an ordered choice model structure with a disturbance term following type I extreme value distribution. In our model system, the proportional hazard function for household *i* terminating the survey participation after the household has been participating in the survey for *u* waves can be described as follows,

$$l_i(u) = l_0(u)\exp(b' x_{iu} + a w_{iu} + g n_{iu} + e_i + x_i + z_i), \qquad (6\text{-}1)$$

where x_{iu} is a vector of household demographics and survey characteristics, and w_{iu} and n_{iu} are the number of home-based work and home-based non-work trips respectively. β, α, and γ are corresponding fixed/random coefficients to be estimated. A total of three disturbance terms are introduced in the equation. ε_i is included to account for the unobserved heterogeneity across households; another disturbance term x_i captures an unobserved common factor between propensities of survey participation and home-based work trips; a similar term z_i represents an unobserved common variation between survey participation and home-based non-work trips. All three disturbance terms follow the normal distribution with a mean of zero. The mean and the variance for the disturbance terms are shown as follows,

$$\begin{aligned} &E(e_i) = E(x_i) = E(z_i) = 0, \\ &V(e_i) = 1,\ V(x_i) = s_1^2,\ V(z_i) = s_2^2, \\ &Cov(e_i, x_i) = Cov(e_i, z_i) = Cov(z_i, x_i) = 0. \end{aligned} \qquad (6\text{-}2)$$

We assume that ε_i follows the standard normal distribution and the variances for two common disturbance terms x_i and z_i are s_1^2 and s_2^2

respectively, where s_1 and s_2 are the parameters need to be estimated. In the hazard function, the variation due to the random coefficient effect captures the observed heterogeneity, while the standard disturbance term e_i characterizes the unobserved heterogeneity independent of the trip-making behavior.

The survival function for household i after staying with the survey for t waves then can be written as,

$$S(t_i) = \exp\left[-\sum_{u=1}^{t_i}\left(l_0(u)\exp(b'x_{iu} + aw_{iu} + gn_{iu} + e_i + x_i + z_i)\right)\right]$$
$$= \exp\left[-\sum_{u=1}^{t_i}\exp(h_u + b'x_{iu} + aw_{iu} + gn_{iu} + e_i + x_i + z_i)\right] \quad (6\text{-}3)$$

where $h_u = \ln\left(l_0(u)\right)$. The probability for household i staying in the survey for exactly k waves can be derived as,

$$\Pr(t_i = k \mid w_{iu}, n_{iu}, e_i, x_i, z_i) = S(k) - S(k+1)$$
$$= \exp\left[-\sum_{u=1}^{k-1}\exp\left(h_u + b'x_{iu} + aw_{iu} + gn_{iu} + e_i + x_i + z_i\right)\right]$$
$$-\exp\left[-\sum_{u=1}^{k}\exp\left(h_u + b'x_{iu} + aw_{iu} + gn_{iu} + e_i + x_i + z_i\right)\right] \quad (6\text{-}4)$$

In ordered probit models for trip frequencies, the propensity for household i making home-based work in wave u can be written as,

$$w_{iu}^{*} = q' y_{iu} + u_i + r_u + d_{iu} + y_i + x_i, \quad w_{iu} = m \ if \ m_{w,m} < w_{iu}^{*} < m_{w,m+1}.$$

Similarly, the propensity for home-based non-work trips is

$$n_{iu}^{*} = j' z_{iu} + v_i + t_u + d_{iu} + y_i + z_i, \quad n_{iu} = m \ if \ m_{n,m} < n_{iu}^{*} < m_{n,m+1}, \tag{6-5}$$

where y_{iu} and z_{iu} are the vectors of household demographics, q' and j' are corresponding fixed/random coefficients to be estimated, and m_w's are the thresholds for the work trips and m_n's are the thresholds for the non-work trips. Each latent propensity includes five disturbance terms to accommodate the heterogeneity bias in the panel data and to reflect the endogenous correlations among households' survey participation and trip making behavior. In the parallel equations, the first three disturbance terms are standard disturbance terms in the two-way random effect model for panel data (Baltagi, 1995), where u_i and v_i denote the unobserved individual specific effect, r_u and t_u accommodate the unobservable time effect, and d_{iu} is the standard disturbance across time and households. For the last two error terms, y_i denotes the unobserved common factor for home-based work and non-work trip-making behavior, same as x_i for survey participation and work trips and z_i for survey participation and non-work trips.

All the disturbance terms are assumed following normal distribution. The mean and variance can be expressed as,

$$E(u_i) = E(v_i) = E(r_u) = E(t_u) = E(d_{iu}) = E(y_i) = 0,$$
$$V(u_i) = s_3^2, \quad V(r_u) = s_4^2,$$

$$V(v_i) = s_5^2, \quad V(t_u) = s_6^2,$$
$$V(d_{iu}) = 1, \quad V(y_i) = s_7^2,$$
$$Cov(u_i, v_i) = Cov(u_i, y_i) = Cov(v_i, y_i) = 0,$$
$$Cov(u_i, d_{iu}) = Cov(y_i, d_{iu}) = Cov(v_i, d_{iu})$$
$$= Cov(r_u, d_{iu}) = Cov(t_u, d_{iu}) = 0,$$
$$Cov(A_i, B_u) = 0, \quad A = u, v, y \text{ and } B = r, t. \qquad (6\text{-}6)$$

Among these disturbance terms, the disturbance term across time and households (i.e., d_{iu}) follows the standard normal distribution with a mean of zero and a variance of one. The variances for other disturbance terms are the parameters that need to be estimated.

Following the ordered probit model structure, the conditional probability for household *i* making *m* home-based work trips in wave *u* can be obtained as,

$$\Pr(w_{iu} = m \mid u_i, r_u, y_i, x_i) = \Phi(m_{w,m+1} - q' y_{iu} - u_i - r_u - y_i - x_i) - \Phi(m_{w,m} - q' y_{iu} - u_i - r_u - y_i - x_i) \qquad (6\text{-}7)$$

Substituting the conditional probabilities of the trip frequencies into equation (6-4), we obtain the conditional probability of a household *i* with survey participation duration of *k* waves in the panel as

$$\Pr(t_i = k \mid e_i, x_i, z_i, y_i, u_i, v_i, r_u, t_u) = \Pr(t_i = k \mid w_{iu}, n_{iu}, e_i, x_i, z_i) \times \prod_{u=t_0}^{k} \Pr(w_{iu} \mid u_i, r_u, y_i, x_i) \times \prod_{u=t_0}^{k} \Pr(n_{iu} \mid v_i, t_u, y_i, z_i) \qquad (6\text{-}8)$$

where t_0 is the first wave when household i entered the survey. The unconditional probability then can be computed by taking the multiple integrals,

$$\Pr(t_i = k) = \int_{e_i}\int_{x_i}\int_{z_i}\int_{y_i}\int_{u_i}\int_{v_i}\int_{r_u}\int_{t_u} \Pr(t_i = k \mid e_i, x_i, z_i, y_i, u_i, v_i, r_u, t_u) \, de_i dx_i dz_i dy_i du_i dv_i dr_u dt_u \tag{6-9}$$

The log-likelihood function then can be written as,

$$LL = \sum_{i=1}^{N}\sum_{k=1}^{K} M_{ik} \times \log[\Pr(t_i = k)] \tag{6-10}$$

where $M_{ik} = 1$ if $t_i = k$ and $M_{ik} = 0$ otherwise, N is the total number of households in the panel survey, and K is the total number of panel waves.

6.4 MODEL ESTIMATION

6.4.1 Monte Carlo and Quasi-Monte Carlo Methods

The log-likelihood function of the model system involves the evaluation of multidimensional integrals. Sometimes the analytical approach can be used to reduce the dimensionality by introducing an instrument variable, especially for low-dimension integrals. However, the model system proposed in this chapter has eight disturbance terms involved in the integration. Some of them are correlated across households, and some others are correlated across waves. It is unlikely to reduce the dimensionality of the integral without running into a black hole of deriving complex formulas.

Another approach is to evaluate the multidimensional integrals numerically using simulation techniques. We adopt this approach for the model estimation. In the past decade, the use of simulation in estimating econometric models has grown rapidly because of its straightforward concept and the fast expanding computation power. The simulation techniques used in model estimation free the researchers to specify models that fit better in the behavioral realism. Among discrete choice models, for example, multinomial logit and nested logit models have been long established for the choice analysis. The limitations of these first-generation models, mainly due to the property of Independence from Irrelevant Alternatives (IIA), are well recognized at the time. To a large extent, the barrier to overcome these limitations lies in the difficulties in model estimation. Tremendous progress has been made over the past decade since McFadden introduced simulation methods in 1989 that make it practical to estimate discrete choice models with more flexible structure. The estimation power relaxes the rigid behavioral restrictions posing on the early models. Recent developments include mixed multinomial logit and integrated choice and latent variable models which are able to accommodate individual heterogeneity, taste variation, and influences of attitude and perceptions on the decision making process. Simulation methods are widely used to estimate these models, particularly in the numerical integration process to obtain estimated probabilities.

The history of numerical integration can be dated back to the invention of calculus partly because the integrals of elementary functions, in general, can not be computed in close form during the 18^{th} and 19^{th} centuries (Press *et al.*, 1992).

The invention of computer led the numerical integration of differential equations to a much richer and more feasible field. Monte Carlo method, for instant, has been broadly applied to many diverse fields, from the simulation of complex physical phenomena to the simulation of games of chance.

The Monte Carlo method is a general expression which depicts a stochastic technique based on the use of random numbers and probability statistics to investigate problems. There are two major Monte Carlo techniques for evaluating definite integrals. The first method is similar to the rejection method of generating random variables for arbitrary distribution functions. Suppose we want to integrate a function g over a region W which has a complicated shape. We may draw points at random uniformly within a bounding box V which includes W and can be easily sampled, then the integral of g over W is,

$$\int_W g(x) = \frac{n^*}{n} V , \qquad (6\text{-}11)$$

where n^* is the number of points within region W, n is the total number of random draws, and V is the volume of the bounding box. This method is very inefficient since many points are required to make the right hand side of the equation (6-11) converge to the left side of the equation.

A more effective approach is to approximate the integral by the application of mean value theorem of calculus. Consider the one dimensional integral,

$$E = \int_a^b f(x) . \qquad (6\text{-}12)$$

The integral can be approximated by

$$E_N = \frac{(b-a)}{N}\sum_{i=1}^{N} f(x_i), \qquad (6\text{-}13)$$

where the points x_i are pseudo-random sequences fully covering the range of integration. In the limit of large number N, E_N tends to the exact value of E. Estimation based on equation (6-13) converges much quicker than those based on equation (6-11). Thus, we adopt a procedure similar to equation (6-13) to evaluate the multidimensional integrals in the log-likelihood function.

6.4.2 Halton Sequence

If pseudo-random numbers are used for the Monte Carlo evaluation of integrals, there are always some regions of the integral that are underrepresented as well as overrepresented due to the clumps and voids in any given sample. Monte Carlo integration inevitably suffers from the flaws in random number generators. In this case, higher accuracy can only be achieved by increasing the number of random draws. In general, more iteration leads the integration estimate to converge towards the actual solution as $1/N^{0.5}$ independently of integral dimension where N is the total number of random samples.

Monte Carlo integration can be efficient in the case of multidimensional integrals compared to other method such as Trapezoidal method. However, in practice the Monte Carlo integration of multivariate functions over a multidimensional region is not always very efficient. This is mainly because the number of random draws needed to sample S-dimensional space increases as the S^{th} power of the number needed for one-dimensional integral to reach the same

level of accuracy. It is a common feature that when a random set of *S*-dimensional space is generated, the resulting distribution probability is either significantly higher or lower than the expected. The solution to this problem is also a large number of repetitions.

In some cases, such as using simulation techniques to describe an arrival process, the large number of repetitions may not be a big problem. However, when the simulation is a part of the optimization of a complex function, the number of simulation iterations is crucial for the convergence speed and estimation accuracy. The estimation of discrete choice models, for instance, involves maximization of some function, such as the likelihood function or the moment conditions. Even though the application of simulation methods has led the discrete choice modeling to a new generation, an efficient random number generator is critical to estimate models with behaviorally realistic structure because the likelihood function is often non-linear in nature and sometimes not well-behaved. Often we need to balance both sides when applying simulation methods to solve optimization problems. On the simulation side, the simulated probability distribution should well represent the expected one. A better coverage general means more repetitions. On the optimization side, more repetitions tremendously increase the computing burden and therefore result in a much slower convergence towards the optimal solution with the uncertainty brought up by the random realizations of the probability distribution. Reducing the uncertainty in the random sequence and fewer random samples certainly will accelerate the solution searching process.

An effective way to reduce the uncertainty in the random samples and have a more evenly scattered coverage is quasi-Monte Carlo methods which use quasi-random numbers instead of pseudo-random numbers for the simulation (Bhat, 2001). Quasi-random sequence sometimes is referred to as a low-discrepancy sequence where the notion of discrepancy is used to quantify the quality of uniformity of a finite point set. It has a more uniform behavior than the pseudo-random sequence. Fewer quasi-random points are needed to reach a similar level of accuracy as obtained by pseudo-random sequence. Quasi-Monte Carlo methods combine the advantages of Monte Carlo and uniform lattice methods. The error bound for the quasi-Monte Carlo methods is in the order of $((\ln N)^S/N)$, where S is the dimension and N is the total number of samples. This error bound suggests a potentially faster convergence than Monte Carlo methods. The convergence rate can be as fast as $1/N$ for reasonable well-behaved smooth functions. Moreover, unlike Monte Carlo methods where the error bound is probabilistic, the quasi-Monte Carlo methods guarantee the accuracy in a deterministic way because of the deterministic nature of quasi-random sequences.

Despite the meaning of "random", quasi-random numbers are highly equidistributed deterministic points. Pseudo-random sequences include the Hammersley, Halton, sobol', Faure, generalized Niederreiter and other sequences (Morokoff and Caflisch, 1994; Niederreiter, 1992; Tezuka, 1995). We use Halton sequences for its conception simplicity (Bhat, 2001). A Halton sequence is defined in terms of a given number, usually a prime number because the sequence based on non-prime number would create clumps with loss of efficiency. The

standard Halton sequence for one dimension corresponding to prime number 3, for example, is

$$\frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{4}{9}, \frac{7}{9}, \frac{2}{9}, \frac{5}{9}, \frac{8}{9}, \frac{1}{27}, \frac{10}{27}, \frac{19}{27}, \frac{4}{27}, \frac{13}{27}, \frac{22}{27}, \ldots\ldots$$

The example shows that the sequence is generated iteratively. First, the unit [0, 1] is divided into three segments with breakpoints at 1/3 and 2/3. Then, each segment is further divided into three segments, and so as each further divided segment. The breakpoints enter the sequence in a particular way. Let R_t denote a series of numbers at iteration t. The algorithm to generate n standard Halton sequence numbers can be described as follows, using prime number p as a base.

1. Initialize the sequence with $t = 0$ and $R_t = \{ 0 \}$;
2. Update t and R_t with $t = t + 1$ and $R_t = \{ R_{t-1}, R_{t-1} + 1/p^t, R_{t-1} + 2/p^t, \ldots, R_{t-1} + (p-1)/p^t \}$;
3. If the total number of sequence in R_t has reached n, stop. Otherwise, go to step 2.

Similarly, the standard Halton sequence in S dimensional domain is generated by paring S one dimensional sequences based on S prime numbers. It should be noted that one issue with Halton sequences arises when they are used for high dimensional integrals. For high dimensional integrals, the sequences need large prime numbers. However, Halton sequences defined by large prime

numbers can be highly correlated with each other over a large portion of random draws. Bratten and Weller (1979) pointed out the existence of this problem in Halton sequence and Bhat (2001) demonstrates the correlation among the standard Halton sequences using prime number 43 and 47 (for 14^{th} and 15^{th} dimension). In this case, the Halton draws fall in a rapid deterioration in the uniformity which leads to a highly correlated structure. In general, the deterioration is clearly noticeable beyond five dimensions (with prime number greater than 13).

The correlation can be broken up to improve the uniformity by "scrambling" the order of the sequence when generating the random number iteratively. There are different ways for the scrambling. Bhat (2001) adopted Bratten and Weller's approach (1979) to permute the order of the random draws. The procedure can be implemented by a slight change to update R_t in step 2 of the algorithm. Let O denote the permutation sequence $O = \{o_1, o_2, o_3, \ldots, o_{p-2}, o_{p-1}\}$, then the updating procedure becomes: $R_t = \{ R_{t-1}, R_{t-1} + o_1/p^t, R_{t-1} + o_2/p^t, \ldots, R_{t-1} + o_{p-2}/p^t, R_{t-1} + o_{p-1}/p^t\}$. The standard Halton can be considered as a special case with $O = \{1, 2, 3, \ldots, p\text{-}2, p\text{-}1\}$. Use prime number 3 as an example. The permutation sequence is $O = \{2, 1\}$ instead of $\{1, 2\}$. Then, the scrambled Halton draws become

$$\frac{2}{3}, \frac{1}{3}, \frac{2}{9}, \frac{8}{9}, \frac{5}{9}, \frac{1}{9}, \frac{7}{9}, \frac{4}{9}, \frac{2}{27}, \frac{20}{27}, \frac{11}{27}, \frac{8}{27}, \frac{26}{27}, \frac{17}{27}, \ldots\ldots$$

A recent computational experiment undertaken by Bhat accesses the performance of the traditional pseudo-random sequence, the standard Halton sequence, and the scrambled Halton sequence in the estimation of mixed logit models. The results indicate that the scrambled Halton sequence outperformed the standard Halton and the pseudo-random sequences in terms of estimation accuracy and the number of draws. Thus, we use the scrambled Halton sequence for our model estimation.

Once random numbers with uniform distribution in the region of [0, 1] is generated, they can be transformed to follow a standard normal distribution. The transformation is made by using the inverse function shown as

$$R_s = \Phi^{-1}(H) \tag{6-14}$$

where R_s is the random realizations of the standard normal distribution, H is the scrambled Halton sequence following uniform distribution, and Φ^{-1} is the inverse function of normal distribution. The normal distribution with a mean of zero and a variance of σ^2 then can be derived from the standard normal,

$$R_s = R_s \times s \tag{6-15}$$

6.4.3 Simulated Likelihood Function

The probability of a household i participating in the panel survey for k waves, as shown in equation (6-9), involves integrals over eight disturbance terms that accommodate individual heterogeneity, random-effects over time, and correlations among trip frequencies and survey participation. Thus, the scrambled Halton sequence is generated using the first eight primes. Let R_e , R_x,, R_z , R_y , R_u

, R_v ,R_r , R_t ,R_s denote the random sequences for e, x, z, y, n, v, r, and t respectively. Each sequence includes H random points and $R_{e,r}$ denotes the r^{th} number in the sequence. The simulated probability then is obtained as the mean of the realizations of the estimated probability based on these random sequences and can be expressed as

$$\tilde{\Pr}(t_i = k) = \frac{1}{H}\sum_{r=1}^{H}\left\{\Pr_r(t_i = k)\times\prod_{u=t_0}^{k}\Pr_r(w_{iu})\times\prod_{u=t_0}^{k}\Pr_r(n_{iu})\right\} \qquad (6\text{-}16)$$

where $\Pr_r$ denotes the probability computed corresponding to the r^{th} number in the sequences, and

$$\Pr_r(t_i = k) = \exp\left[-\sum_{u=1}^{k}\exp\left(h_u + b'x_{iu} + aw_{iu} + gn_{iu} + R_{e,r} + R_{x,r} + R_{z,r}\right)\right] - \exp\left[-\sum_{u=1}^{k+1}\exp\left(h_u + b'x_{iu} + aw_{iu} + gn_{iu} + R_{e,r} + R_{x,r} + R_{z,r}\right)\right], \qquad (6\text{-}17)$$

$$\Pr_r(w_{iu} = m) = \Phi(m_{w,m+1} - q'y_{iu} - R_{u,r} - R_{r,r} - R_{y,r} - R_{x,r}) - \Phi(m_{w,m} - q'y_{iu} - R_{u,r} - R_{r,r} - R_{y,r} - R_{x,r}), \qquad (6\text{-}18)$$

$$\Pr_r(n_{iu} = m) = \Phi(m_{n,m+1} - j'z_{iu} - R_{v,r} - R_{t,r} - R_{y,r} - R_{z,r}) - \Phi(m_{n,m} - q'y_{iu} - R_{v,r} - R_{t,r} - R_{y,r} - R_{z,r}). \qquad (6\text{-}19)$$

The data used model estimation consists of 4802 households who have at least participated in one wave of the PSTP survey. The model is estimated using MAXLIK module implemented in the econometric package GAUSS. The

analytical gradient function for the parameters is also coded to achieve a faster convergence.

6.5 EMPIRICAL RESULTS

The estimated coefficients of the joint model system are shown in Table 6-1, 6-2, and 6-3 for survey participation, home-based work trips, and home-based non-work trips respectively. To compare the estimation results, we also develop models for home-based work and non-work trips without consideration of selectivity bias correction. One set of models adopt a random-coefficient ordered probit model structure to accommodate heterogeneity across households in the panel data. The results are shown in Table 6-4 for home-based work and non-work trips and Table 6-6 for home-based non-work trips. The other set of models use the standard ordered response probit structure. The results are presented in Table 6-5 and 6-7.

6.5.1 Survey Participation

The estimated coefficients of eternal covariate effects on survey participation duration are presented in Table 6-1. Compared with the early model results shown in Table 5-3 and 5-4, the sign of the coefficients remain the same. Once again, the early models experience a pattern of toward-zero bias. The absolute values of the coefficients in Table 6-1 are greater than those in Table 5-3 and 5-4.

6.5.1.1 Household Demographic Variables

The estimation results show that household life cycle plays an important role in survey participation duration. Compared to households with all children under 5 years old, households with children at the age of 6 to 17 stay longer with the survey. It is again observed that households with young adults (younger than 35) and no children are more likely to terminate their participation in the survey. Among these households, the single-adult households are more likely to decline the survey request, followed by households with two or more adults. The covariate effects also indicate that households with no children and older adult members tend to respond to the survey for more waves, especially for households with adults older than 65. It is also found that split households and households with income between 25k and 45k are more likely to terminate their survey participation than others.

Another important demographic variable is the number of workers in household. The results show that the more workers in the household, the more likely it will continue responding to the panel surveys. Furthermore, the statistics show that part of the variation across households can be accommodated by the random effect of this variable. The observed and unobserved heterogeneity are discussed in detail in Section 6.5.1.3.

6.5.1.2 Survey-Related Attributes and Trip Frequency

The estimated coefficients of survey-related attributes suggest that survey burden, in general, is likely to lead to survey nonresponse. For instance, the

model results show that households with more members eligible to fill out the travel diary are more likely to stop responding to the survey. The item-nonresponse indicators also provide some insights on survey participation duration. An item nonresponse of household demographic variable (except income information) implies that the household is reluctant to provide the information and therefore is more likely to subject to nonresponse in the next wave. However, a missing value in income does not indicate the same trends. The estimated coefficient shows that households with no income information provided tend to continue participate in the survey for the following wave. The results also indicate that households in regular carpooling sample group are less likely to continue their survey participation.

Considered as an indicator of survey burden, the number of home-based work trips has a positive impact on the hazard function, indicating that households with more work trips tend to terminate their survey participation in the next wave. On the other hand, the home-based non-work trips are negatively associated with the hazard function probably due to the joint effect of survey burden and time constraint.

6.5.1.3 The Observed and Unobserved Heterogeneity

In the model developed earlier in this dissertation (see Chapter 5), heterogeneity among individual households is captured by a Gamma distributed disturbance term with a variance parameter that need to be estimated. In the joint model system heterogeneity is accommodated by three segments: a disturbance

term following a standard Gamma distribution, common disturbance terms among participation duration and trip frequency models, and variation in the random covariate effects. The first two segments capture the unobserved heterogeneity and the last segment characterizes the observed heterogeneity.

We tried to randomize the coefficients of eternal covariates to see how much of the heterogeneity can be distinguished by household demographic attributes and survey characteristics. The final specification includes two randomized coefficients for the number of home-based non-work trips and the number of workers respectively. The random coefficient for the number of home-based non-work trips has a mean of -0.0449 and a standard deviation of 0.0149. The ratio of mean over standard deviation is about one third for this variable. The coefficient for the number of workers has a mean of -1.5143 and a standard deviation of 0.9832 with a ratio of mean over standard deviation of 0.65.

It is not too difficult to outline the reasons for the variation existing in the covariate effects of the home-based non-work trips and household workers. These two variables reflect counterpart impacts on the survey participation decision. On one hand, workers have more restricted time constraints and more home-based non-work trips increase the burden to fill out the travel diary. This side of the effects accelerates a household's pace to quit the survey response. On the other hand, workers are more concerned about traffic problems and have more regular travel schedule during weekdays; and more home-based non-work trips also indicate that less restricted time constraints are imposed on household members. This side of the effects actually makes individual household more

committed to the survey. The mean values of the estimated coefficients illustrate the combined consequences observed in the data. A negative mean value for the home-based non-work trips, for instance, indicates that the effect of less time constraints overcomes the effect of survey burden. The standard deviations then suggest that the variation due to the counterpart effects can not be ignored.

We found that common disturbance terms across the duration model and trip frequency models are not significant. Summary statistics for these variables are still presented in Table 6-1. The insignificant estimates indicate that, in terms of survey participation duration, heterogeneity across households can fully be accommodated by observed factors. Furthermore, we should emphasize that these insignificant estimates do not imply the lack of endogenous correlation between survey participation duration and trip frequency, especially between survey participation and home-based non-work trips, because of the random effect of home-based non-work trips on the hazard function. The results indicate that the endogenous correlation can be captured by observed factors in stead of unobserved factors.

6.5.2 Home-Based Work Trips

The model results for home-based work trips are presented in Table 6-2, 6-4, and 6-5. The coefficients in Table 6-2 are estimated in the joint model system. The coefficients in Table 6-4 are estimated using a random-coefficient ordered probit model structure without consideration for selectivity bias correction, and same as those in Table 6-5 which are estimated using standard

ordered probit model structure. In these three models, the estimated coefficients have the same sign but with different values.

In all three models, coefficients for dummy variables representing panel waves are positive. In the joint model and the standard ordered probit model, the coefficients are positive and significant, although the coefficients are not significant in the random-coefficient ordered model. The positive sign suggest that households make more home-based work trips in later waves than in wave 1, an increase across time. In addition, the model results indicate that households with older adults, with children, and in low-income group make fewer trips than others; and households with more workers and adults make more work trips.

Results in the joint model system and random-coefficient model indicate that heterogeneity across households is significant. In both models, the variation across households' work-trip-making behavior can be captured by the random effect of the number of children at the age of 6 to 17. It is probably due to the variation in woman-in-work-force for households with children in school. Meanwhile, the unobserved heterogeneity accommodated by a disturbance term across households is found significant in the joint model with a standard deviation of 0.6508.

6.5.3 Home-Based Non-Work Trips

The estimated coefficients for home-based non-work trips are presented in Table 6-3, 6-6, and 6-7, corresponding to the joint model system, random-coefficient ordered probit model, and standard ordered probit model respectively.

Our model results show that the estimated coefficients remain the same sign but with different magnitudes in different models. In the joint model and the random-coefficient model, heterogeneity across time is again fully captured by the fixed dummy variables and the random effect across waves is not significant. All the dummy variables for later waves have a negative sign, suggesting that households make fewer trips in later waves than in wave 1. The negative coefficients may reflect households' travel pattern across waves, or could be a result of under-reported travel diary.

The models indicate that households with children make more home-based non-work trips than those with no children, especially for households with children at age 6 to 17. In addition, households with more workers make fewer non-work trips and households in higher income group make more non-work trips.

The model results also suggest that the observed heterogeneity across households is mainly due to the random effects of the number of adults and the number of children at age 6 to 17. It should be noted that the estimated standard deviation over the estimated mean for the number of children at age 6 to 17 is 1.77 (0.6421/0.3621), indicating that a large variation in home-based non-work trip-making behavior is associated with teen-agers.

In summary, Households with more adults, more teen-agers, and higher income tend to make more home-based non-work trips. Households with more workers are more likely to make fewer home-based non-work trips. The results are consistent with findings in other studies.

6.6 SUMMARY

This chapter proposes a multi-level modeling system to estimate survey participation duration and trip frequencies for various trip purposes. The model structure accommodates heterogeneity across time and individuals. Meanwhile, both exogenous and endogenous correlations among survey participation duration and trip frequencies are reflected in the model structure.

The estimation results suggest that there are endogenous correlations among the survey participation duration and home-based non-work trip rates. Ignoring this endogenous correlation leads to biased estimates. The results also show that heterogeneity across households can be fully captured by the random effect of certain independent variables, such as the number of workers for the duration model and the number of children at age 6 to 17 for trip frequency models.

In terms of households' trip-making behavior, we found that households with more adults and more workers make more home-based work trips and households with senior citizens make fewer work trips. For home-based non-work trip, it is found that households with higher income, older adults, and fewer workers make more non-work trips. These results are consistent with findings of other studies.

The next chapter summarizes the findings in this dissertation, as well as provides recommendations and topics of future research.

Table 6-1: Survey participation duration in joint model system

Independent Variable	Coefficient	t-value	Significant Level
Dummy variable for households entering the panel in wave 2	1.0784	9.331	0
Dummy variable for households entering the panel in wave 3	0.7642	6.416	0
Dummy variable for households entering the panel in wave 4	1.2256	11.901	0
Dummy variable for households entering the panel in wave 5	1.5256	11.197	0
Dummy variable for households entering the panel in wave 6	1.8743	13.724	0
Household life cycle			
Households with children between the age of 6 to 17	-0.3265	-3.191	0.0014
Households with no children and one adult under the age of 35	0.4806	3.014	0.0026
Households with no children and one adult between the age of 36 to 64	-0.7286	-5.370	0
Households with no children and one adult older than 65	-1.3819	-8.611	0
Households with no children and two+ adults under the age of 35	0.3554	2.530	0.0014
Households with no children and two+ adults between the age of 36 to 64	-0.6584	-6.405	0
Households with no children and two+ adults older than 65	-1.9327	-15.429	0
Other household demographics			
Household income: 25k to 45k	0.2227	2.891	0.0019
Split household	0.4845	2.262	0.0237
Number of workers	-1.5147	-26.011	0
Standard deviation	0.9832	12.955	0

Table 6-1: Survey participation duration in joint model system (cont.)

Independent Variable	Coefficient	t-value	Significant Level
Survey-Related Attributes and Trip Frequencies			
Number of household members who filled out the travel diary	0.4037	6.960	0
Household demographic attributes (other than income) are imputed	0.3872	5.508	0
Household income is imputed	-0.5139	-5.810	0
Number of home-based work trips	0.068	5.738	0
Number of home-based non-work trips	-0.0449	-7.437	0
Standard deviation	0.0149	-2.011	0.0443
Disturbance Term			
Common disturbance term for the duration and the HBW trip model			
Standard deviation	0.0079	0.126	0.8995
Common disturbance term for the duration and HBNW trip model			
Standard deviation	0.0589	1.461	0.1441
Common disturbance term for HBW and HBNW trip model			
Standard deviation	0.0115	0.626	0.5314

Table 6-2: Model for home-based work trips

Independent variable	Coefficient	t-value	Significant level
Dummy variable for wave 2	0.2245	6.172	0
Dummy variable for wave 3	0.1754	4.637	0
Dummy variable for wave 4	0.1981	5.275	0.0032
Dummy variable for wave 5	0.1141	2.952	0.01
Dummy variable for wave 6	0.1021	2.578	0
Dummy variable for wave 7	0.2294	5.785	0
Household Life Cycle			
Household with children between the age of 6 to 17	0.3442	7.283	0
Households with no children and one adult under the age of 35	0.6725	9.107	0
Households with no children and one adult between the age of 36 to 64	0.3276	5.480	0
Households with no children and one adult older than 65	-0.8860	-10.676	0
Households with no children and two+ adults under the age of 35	0.8695	15.482	0
Households with no children and two+ adults between the age of 36 to 64	0.2601	6.166	0
Households with no children and two+ adults older than 65	-1.0447	-18.724	0
Other household demographics			
Number of workers	0.585	35.507	0
Number of adults	0.6378	23.098	0
Household income: < 25K	-0.1655	-5.087	0
Number of children between the age of 6 to 17 Standard deviation	0.1359	5.183	0
Disturbance term			
Unobserved heterogeneity across household: standard deviation	0.6508	33.162	0

Table 6-3: Model for home-based non-work trips

Independent variable	Coefficient	t-value	Significant level
Dummy variable for wave 2	-0.1733	-5.001	0
Dummy variable for wave 3	-0.2247	-6.222	0
Dummy variable for wave 4	-0.2573	-7.144	0
Dummy variable for wave 5	-0.3019	-8.147	0
Dummy variable for wave 6	-0.2557	-6.698	0
Dummy variable for wave 7	-0.2987	-7.748	0
Household Life Cycle			
Households with no children and one adult under the age of 35	-0.8579	-12.159	0
Households with no children and one adult between the age of 36 to 64	-0.7587	-12.961	0
Households with no children and one adult older than 65	-0.3962	-5.914	0
Households with no children and two+ adults under the age of 35	-0.7133	-12.022	0
Households with no children and two+ adults between the age of 36 to 64	-0.5276	-11.711	0
Other household demographics			
Number of workers	-0.1645	-9.725	0
Household income: 25K to 45K	0.0860	2.742	0.0061
Household income: 45K to 75K	0.1466	4.131	0
Household income: > 75K	0.1418	3.008	0.0026
Number of adults	0.7619	22.635	0
Standard deviation	0.3542	30.921	0
Number of children between the age of 6 to 17	0.2644	8.328	0
Standard deviation	0.3915	18.137	0
Disturbance term			
Unobserved heterogeneity across household: standard deviation	0.1671	2.622	0.0087

Table 6-4: Random-coefficient ordered response probit model for HBW trips (without accommodating selectivity bias)

Independent Variable	Coefficient	t-value	Significant level
Dummy variable for wave 2	0.3016	0.775	0.4382
Dummy variable for wave 3	0.2243	1.128	0.2594
Dummy variable for wave 4	0.1546	0.811	0.4174
Dummy variable for wave 5	0.1669	0.587	0.5574
Dummy variable for wave 6	0.1205	1.082	0.2793
Dummy variable for wave 7	0.296	0.52	0.6033
Household Life Cycle			
Households with children between the age of 6 to 17	0.3561	9.894	0
Households with no children and one adult under the age of 35	0.6401	10.068	0
Households with no children and one adult between the age of 36 to 64	0.3859	8.202	0
Households with no children and one adult older than 65	-0.6602	-9.509	0
Households with no children and two+ adults under the age of 35	0.7951	14.08	0
Households with no children and two+ adults between the age of 36 to 64	0.2564	7.827	0
Households with no children and two+ adults older than 65	-0.928	-16.159	0
Other Household Demographics			
Number of workers	0.6112	20.102	0
Number of adults	0.5658	16.924	0
Household income: < 25K	-0.1294	-4.665	0
Number of children between the age of 6 to 17			
Standard deviation	0.1418	5.000	0

Table 6-5: Standard ordered response probit model for HBW trips (without accommodating selectivity bias)

Independent Variable	Coefficient	t-value	Significant level
Dummy variable for wave 2	0.2222	6.340	0
Dummy variable for wave 3	0.1817	5.082	0
Dummy variable for wave 4	0.1836	5.259	0
Dummy variable for wave 5	0.1098	3.076	0.0021
Dummy variable for wave 6	0.0989	2.771	0.0056
Dummy variable for wave 7	0.1830	5.162	0
Household Life Cycle			
Households with children between the age of 6 to 17	0.3401	11.028	0
Households with no children and one adult under the age of 35	0.6156	11.080	0
Households with no children and one adult between the age of 36 to 64	0.3693	8.671	0
Households with no children and one adult older than 65	-0.6497	-10.641	0
Households with no children and two+ adults under the age of 35	0.7651	17.791	0
Households with no children and two+ adults between the age of 36 to 64	0.2444	8.268	0
Households with no children and two+ adults older than 65	-0.9058	-23.024	0
Other Household Demographics			
Number of workers	0.5883	42.23	0
Number of adults	0.5471	24.823	0
Household income: < 25K	-0.1252	-4.759	0

Table 6-6: Random-coefficient ordered response probit model for HBNW trips (without accommodating selectivity bias)

Independent Variable	Coefficient	t-value	Significant level
Dummy variable for wave 2	-0.8118	-2.085	0.0371
Dummy variable for wave 3	-0.5731	-2.86	0.0042
Dummy variable for wave 4	-0.0594	-0.301	0.7637
Dummy variable for wave 5	-0.7819	-2.73	0.0063
Dummy variable for wave 6	-0.4751	-4.072	0
Dummy variable for wave 7	-1.1813	-2.07	0.0384
Household Life Cycle			
Households with children between the age of 6 to 17	0.1875	2.897	0.0038
Households with no children and one adult under the age of 35	-0.8141	-10.287	0
Households with no children and one adult between the age of 36 to 64	-0.7155	-10.793	0
Households with no children and one adult older than 65	-0.3079	-4.22	0
Households with no children and two+ adults under the age of 35	-0.9463	-12.615	0
Households with no children and two+ adults between the age of 36 to 64	-0.6414	-11.83	0
Other Household Demographics			
Number of workers	-0.296	-12.426	0
Household income: 25K-45K	0.1508	4.077	0
Household income: 45K-75K	0.2383	5.688	0
Household income: >75K	0.2636	4.786	0
Number of adults	1.2407	20.649	0
Standard deviation	0.591	19.222	0
Number of children between the age of 6 to 17	0.3632	9.362	0
Standard deviation	0.6427	17.623	0

Table 6-7: Ordered response probit model for HBNW trips (without accommodating selectivity bias)

Independent Variable	Coefficient	t-value	Significant level
Dummy variable for wave 2	-0.1667	-4.977	0
Dummy variable for wave 3	-0.1950	-5.729	0
Dummy variable for wave 4	-0.2255	-6.781	0
Dummy variable for wave 5	-0.2530	-7.463	0
Dummy variable for wave 6	-0.2124	-6.266	0
Dummy variable for wave 7	-0.2182	-6.486	0
Household Life Cycle			
Households with children between the age of 6 to 17	0.1136	3.218	0
Households with no children and one adult under the age of 35	-0.6517	-11.484	0
Households with no children and one adult between the age of 36 to 64	-0.5907	-13.549	0
Households with no children and one adult older than 65	-0.2748	-5.364	0
Households with no children and two+ adults under the age of 35	-0.5907	-13.445	0
Households with no children and two+ adults between the age of 36 to 64	-0.3859	-12.407	0
Households with no children and two+ adults older than 65	0.0651	1.717	0.0859
Other Household Demographics			
Number of workers	-0.1612	30.919	0
Household income: 25K-45K	0.1086	11.225	0
Household income: > 45K	0.1793	-12.016	0
Number of adults	0.6573	4.160	0
Number of children between the age of 6 to 17	0.2016	6.590	0

Chapter 7 Conclusions

7.1 CONTRIBUTIONS

Nonresponse is a classic topic for survey researchers and its association with the subject of interest in survey always attracts attention from modelers. Meanwhile, the continuing growth of computational power and the advances in numeric methods make it possible to test more realistic behavioral hypotheses. This dissertation locates the problems in multi-wave household travel surveys, a first attempt in transportation literature to the author's knowledge. The methodological contributions of this dissertation are as follows:

- Proposed a hazard-based duration model to capture duration dependence in panel survey participation behavior;
- Accommodated lagged impact of exogenous variables on current participation decision;
- Introduced multiple indicators in the model specification to test the behavioral hypothesis between survey participation and travel activity;
- Applied a more efficient quasi-Monte Carlo simulation method for the model estimation.

This dissertation intends to provide quantitative support for a better understanding of nonresponse in longitudinal household travel surveys by taking a disaggregated point of view on sample units. The hypothesis of whether nonresponse is ignorable to travel activity analysis (i.e. trip frequency) is tested as

well. A comprehensive analysis is the goal of this work. The following issues are specifically addressed:

- Impact of survey burden on the decision of responding to the survey
- Impact of other survey feature indicators, such as indicators for a missing item, different sampling group, or another survey conducted at the same time
- Impact of sample unit's demographic characteristics, especially employment status and age group.

7.2 SUMMARY OF FINDINGS

The primary objective of this dissertation is to understand survey participation decision in longitudinal household travel surveys, especially to investigate the impact of survey burden. The dissertation views households' participation decision in panel surveys as a duration process. The duration process utilizes the maximum volume of information gathered in panel surveys and considers the repeated survey participation decisions simultaneously. The duration approach also has advantages of capturing duration dependence and incorporating time-varying covariate effects.

The existing literature has shown that the survey participation decision is affected by social environment, household demographics, and survey features. Our modeling results support this statement. In the empirical analysis using the PSTP data, we found that household demographics account for a large portion of the participation decision. Among household demographics, we found that the

employment status of household members is a major determinant of the survey participation decision. On the surface, it may appear that more workers in the household would lead to a lower response rate due to the time constraint imposed on workers. However, our results indicate that households with more workers stay in the survey for more waves, probably because workers are more concerned about the traffic problems that they encounter frequently. Household life cycle type is another important factor. It is found that households with younger adults are the least likely to respond to the survey. We also found that older households are more likely to participate in the survey than those households with children.

Besides household demographics, our model specification focuses on evaluating the effects of survey burden on survey participation. Different measures of survey burden are incorporated in the model. For instance, the number of household members who are eligible to fill out the travel diary is used to represent the internal cooperation within a household and trip frequencies reported in travel diaries are used to represent the work load to finish the survey. We found that, in general, survey burden has a negative impact on survey participation. However, there is an interesting finding about home-based non-work trip rates. It seems that the more home-based non-work trips a household makes, the more likely this household is to continue participating in the survey. In this case, the survey burden seems to be positively associated with the survey participation. The possible reason is that the more home-based non-work trips during weekdays also reflect a less restricted time constraint imposed on the

household and, therefore, the household is more likely to continue its participation in the survey.

Some survey-related variables are also found significant in the models. Item nonresponse is a good indicator for unit nonresponse in the following waves. A missing value in household demographic variables (except income variable) suggests that this household has much higher likelihood to quit its survey participation in the next wave. In addition, the model results show that a missing value in income variable does not reflect the same indication about households' survey participation as the other missing values. It does not necessarily suggest that the household is more likely to stop responding to the survey. The sample households in the PSTP survey are choice-based sample. We found that households in regular carpooling group have slightly higher probability of terminating their survey participation than those in SOV and transit-user groups.

Another objective of the dissertation is to examine the relationship between survey participation and trip frequency. A modeling system is proposed to estimate the survey participation duration and trip frequencies simultaneously. In the duration model, trip frequencies appear on the right side of the equation as independent variables. Furthermore, the disturbance terms are structured to accommodate the observed and unobserved heterogeneity between them. The model results show that there is an endogenous correlation between the survey participation duration and trip rates for home-based non-work trips. This endogenous correlation cannot be overlooked when modeling the survey participation duration. A toward-zero bias is found for the external covariates

effect when the heterogeneity is ignored in the duration model. In addition, the ignorance of this endogenous correlation also leads to biased estimates for the trip frequency models.

7.3 APPLICATIONS, RECOMMENDATIONS, AND FUTURE RESEARCH

7.3.1 Applications

The findings in this dissertation provide insights for the design of more effectively panel surveys. For instance, the duration dependence revealed by baseline hazard rates shows that there are two sudden increases in the likelihood of terminating survey participation. One occurs after households have been in the survey for two waves. The other occurs after households have been in the survey for five waves. In addition, the results indicate that, after remaining in the survey for five waves, it is almost for sure that a household will quit the survey. The duration dependence can help survey operators understand the longitudinal dynamics across waves and therefore, adopt appropriate survey strategies in advance. Since the hazard rate is higher after two waves, survey operators may place more funds for the third wave. Because a much higher termination probability is observed after five waves, survey operators may consider alternative sampling methods under a budget constraint. For instance, they can rotate sample units every five waves in order to achieve a higher response rate and reduce survey cost. Or, a smaller sample size may be used after two waves and five waves.

A common approach to obtaining higher response rates is to use reminders. This process can be more efficiently conducted if potential nonrespondents are effectively targeted. The model results summarized in previous section can be applied to segment households with higher probability of nonresponse. The segmentation is mainly based on key household demographic variables and survey attributes. In addition, the survive probability derived from the model can be used to develop a weighting mechanism to obtain consistent estimates of other travel behavior indicators.

7.3.2 Recommendations for Survey Design

Based on the work presented in this dissertation, the key recommendations for effective survey strategies to achieve a higher response rate are:

- Attract potential survey participants' commitment to the survey

The significant impact of workers indicate that, once the potential survey participants realize that the survey is beneficial to them, they will be more committed to the survey participation. Therefore, it is important to explain to the potential survey participants how the survey is helpful to improve the quality of their lives in order to attract them to participate in the survey.

- Reduce survey burden

Our analysis indicates that, in general, survey burden in household travel survey has negative impact on the survey participation duration. Consequently, reducing survey burden can increase the response rate. The reduction in survey

burden may be accomplished using GPS devices or through a carefully organized survey questionnaire.

7.3.3 Recommendations for Initial Nonresponse Study and Future Research

The modeling effort in this work is conditional on households' initial responses to the survey. The analytical approach can be extended to examine the initial nonresponse behavior if the information is available to the study. The lack of initial nonresponse analyses in the existing literature is mainly due to the challenge of collecting information on initial nonresponse. When potential survey participants refuse to respond to the survey, it is more difficult to collect information on why they refuse to participate. Given the difficulties in data collection procedure, it is very important to identify the key causes of survey nonresponse in order to ask the right questions at the right time. Based on this study, sources for survey nonresponse can be classified into the following categories:

- Key household demographic attributes;
- Key survey design characteristics;
- Survey burden;
- Lack of concerns to the survey subject;
- Restricted time constraint;
- Interactions of each household member's individual decision and a joint household decision;
- Failure to locate the potential participants.

Consequently, information in these categories should be primarily collected for the initial nonresponse study. With information available, the initial survey participation decision can be included in the analysis.

Another future research topic is to further examine the relationship between survey participation duration and participants' travel attitude. Travel attitude can reflect survey participants' concern to the subject of survey. Thus, including the attitude indicators in the model may improve the accuracy and strengthen the explanatory power of the model.

References

Arentze, T. *et al.* (2000). Determinants of Attrition Rates in Two-Wave, Two-Day Household Activity Diary. *Transportation Research Record*, Vol. 1719, pp. 159-164.

Armoogum, J. and J. Madre (1998). Weighting or Imputations? The Example of Nonresponses for Daily Trips in the French NPTS. *Journal of Transportation and Statistics*, Vol. 1, Issue 3, pp. 53-63.

Atrostic, B. K., N. Bates, G. Burt, A. Silberstein, and F. Winters (1999). Nonresponse in Federal Household Surveys: New Measures and New Insights. Paper Presented at the International Conference on Survey Nonresponse, Portland, OR.

Baltagi, B. H. (1995). *Econometric Analysis of Panel Data*. Chichester: John Wiley and Sons.

Ben-Akiva, M. and S. Lerman (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.

Ben-Akiva, M., *et al.* (2001). Hybrid Choice Models: Progress and Challenges. *Marketing Letters*, forthcoming. [Online] Available: http://www.ce.utexas.edu/prof/bhat/ABSTRACTS/Asilomar.pdf [June 30th, 2002].

Bhat, C. R. (1994). Imputing a Continuous Income Variable from Grouped and Missing Income Observations. *Economics Letters*, Vol. 46, pp. 311-319.

Bhat, C. R. (1996a). A Hazard-Based Duration Model of Shopping Activity with Nonparametric Baseline Specification and Nonparametric Control for Unobserved Heterogeneity. *Transportation Research*, Part B, Vol. 30, Issue 3, pp. 189-207.

Bhat, C. R. (1996b). A Generalized Multiple Durations Proportional Hazard Model with an Application to Activity Behavior During the Work-to-Home Commute. *Transportation Research*, Part B, Vol. 30, Issue 6, pp. 465-480.

Bhat, C. R. (1997). Recent Methodological Advances Relevant to Activity and Travel Behavior Analysis. Resource paper prepared for the IATBR Conference, Austin, Texas.

Bhat, C.R. and V. A. Pulugurta (1998). Comparison of Two Alternative Behavioral Mechanisms for Car Ownership Decisions, *Transportation Research*, Part B, Vol. 32, Issue 1, pp. 61-75.

Bhat, C. R. (2000). Duration Modeling: A Methodological Review with Applications in Activity-Travel Analysis. Working paper, Department of Civil Engineering, The University of Texas at Austin.

Bhat, C. R. and J. L. Steed (2001). A continuous-Time Model of Departure Time Choice for Urban Shopping Trips. Paper presented at the 80th annual meeting of the Transportation Research Board, Washington, D. C.

Bhat, C. R. (2001). Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model. *Transportation Research*, Part B, Vol. 35, Issue 7, pp. 677-693.

Bhat, C. R. and H. Zhao (2002). The Spatial Analysis of Activity Stop Generation. *Transportation Research*, Part B, Vol. 36, Issue 6, pp. 557-575.

Bhat, C. R. (2002). Simulation Estimation of Mixed Discrete Choice Models Using Randomized and Scrambled Halton Sequences. *Transportation Research*, forthcoming.

Bhat, C. R., *et al.* (2002a). Intershopping Duration: An Analysis Using Multiweek Data. Paper presented at the 81st annual meeting of the Transportation Research Board, Washington, D. C.

Bhat, C. R., *et al.* (2002b). An Analysis of the Impact of Information and Communication Technologies on Non-Maintenance Shopping Activities. Working paper, Department of Civil Engineering, The University of Texas at Austin.

Bogen, K. (1996). The Effect of Questionnaire Length of Response Rates—A Review of the Literature. In *Proceedings of the American Statistical Association Section on Survey Research Methods*, pp. 1020-1025.

Botman, S.L. and O.T. Thornberry (1992), Survey Design Features Correlates of Nonresponse, ASA Proceedings of the Section on Survey Research Methods, pp. 309-314.

Bratten, E. and G. Weller (1979). An improved low-discrepancy sequence for multidimensional quasi-Monte Carlo integration, *Journal of Computational Physics*, Vol. 33, pp. 249-258.

Brownstone, D. and X. Chu (1997). Multiply-imputed sampling weights for consistent inference with panel attrition. In T. F. Golob *et al.* (*eds.*) *Panels for Transportation Planning: Methods and Applications*. Boston: Kluwer Academic Publishers. pp. 258-273.

Bye, B. V. and E. S. Schechter (1986). A Latent Markov Model Approach to the Estimation of Response Errors in Multiwave Panel Data. *Journal of the American Statistical Association*, Vol. 81, Issue 294, pp. 375-380.

Chamberlain, G. (1984). Panel Data. In Z. Griliches and M.D. Intriligator (*eds.*) *Handbook of Econometrics*, Vol. 2. New York: Elsevier Science, pp. 1248 –1318.

Chung, J. and K. G. Goulias (1995). Sample selection bias with multiple selection rules: application with residential relocation, attrition, and activity participation in Puget Sound Transportation Panel. *Transportation Research Record*, Vol. 1493, pp. 128-135.

Church, A. H. (1993). Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis. *Public Opinion Quarterly*, Vol. 57, Issue 1, pp. 62-79.

Cosslett, S. R. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica*, Vol. 49, Issue 5, pp. 1289-1316.

Couper, M. P. and R. M. Groves (1996). Social environmental impacts on survey cooperation. *Quality and Quantity*, Vol. 30, Issue 2, pp. 173-188.

De Heer, W. F. and G. Moritz (1997). Data Quality Problems in Travel Surveys: An International Overview. In *Transport Surveys: Raising the Standard (Proceedings of an International Conference on Transport Survey Quality and Innovation).* Grainau, Germany.

Domencich, T. and D. McFadden (1975). *Urban Travel Demand*. [Online] Available: http://emlab.berkeley.edu/users/mcfadden/travel.html [June 30th, 2002].

Falaris, E. M. and H. E. Peters (1998). Survey Attrition and Schooling Choices. *The Journal of Human Resources*, Vol. 33, Issue 2, pp. 531-554.

Freeman, R. B. and J. L. Medoff. (1981) The Impact of the Percentage Organized on Union and Nonunion Wages. *The Review of Economics and Statistics*. Vol. 63, Issue 4, pp. 561-572.

Golob, T. F. and L. van Wissen (1989). A Joint Household Travel Distance Generation and Car Ownership Model. *Transportation Research*, Part B, Vol. 23, Issue 6, pp. 471-491.

Golob, T.F. (1990). The dynamics of travel time expenditures and car ownership decisions. *Transportation Research*, Part A, Vol. 24, pp. 443-463.

Goodwin, P. B. (1997). Have panel surveys told us anything new? In T. Golob, R. Kitamura, and L. Long (eds.), *Panels for Transportation Planning: Methods and Applications.* Kluwer Norwell, MA: Academic Publishers.

Grammig, J. and K. Maurer (2000). Non-Monotonic Hazard Functions and the Autoregressive Conditional Duration Model. *Econometrics Journal*, Vol. 3, pp. 16-38.

Greene, W. H. (1981). Sample selection bias as a specification error: comment. *Econometrica*, Vol. 49, Issue 3, pp. 795-798.

Groves, R. M., R. B. Cialdini and M. P. Couper (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, Vol. 56, Issue 4, pp. 475-495.

Groves, R. M., and M. P. Couper (1998). *Nonresponse in household interview surveys.* New York: Wiley.

Ham, J. C. and R. J. Lalonde (1996). The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training. *Econometrica*, Vol. 64, Issue 1, pp. 175-205.

Han, A. and J. A. Hausman (1990). Flexible Parametric Estimation of Duration and Competing Risk Models. *Journal of Applied Econometrics*, Vol. 5, Issue 1, pp. 1-28.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, Vol. 46, Issue 6, pp. 1251-1271.

Hausman, J. A. and D. A. Wise (1979). Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment. *Econometrica*, Vol. 47, Issue 2, pp. 455-474.

Hausman, J. A. and W. E. Taylor (1981). Panel Data and Unobservable Individual Effects. *Econometrica*, Vol. 49, Issue 6, pp. 1377-1398.

Heberlein, T. A. and R. Baumgartner (1978). Factors Affecting Response Rates to Mailed Questionnaires: A Quantitative Analysis of the Published Literature. *American Sociological Review*, Vol. 43, Issue 4, pp. 447-462.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, Vol. 47, Issue 1, pp. 153-161.

Heckman, J.J. (1981). Statistical Models for Discrete Panel Data. In C. Manski and D. McFadden (*eds*.) *Structural Analysis of Discrete Data*. Cambridge, MA: MIT press, pp. 114-178.

Heckman, J. and B. Singer (1984). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica*, Vol. 52, Issue 2, pp. 271-320.

Helsen, K. and D. C. Schmittlein (1993). Analyzing Duration Times in Marketing: Evidence for the Effectiveness of Hazard Rate Models, Vol. 11, Issue 4, pp. 395-414.

Hensher, D. A. (1987). Issues in the pre-analysis of panel data. *Transportation Research*, Part A, Vol. 21, Issue 4-5, pp. 265-285.

Hensher, D. A. and N. C. Smith (1990). Estimating Automobile Utilisation with Panel Data: An Investigation of Alternative Assumptions for the Initial Conditions and Error Covariances. *Transportation Research*, Part A, Vol. 24, Issue 6, pp. 417-426.

Hensher, D. A., N. C. Smith, F. W. Milthorpe and P.O. Marnard (1992). *Dimensions of Automobile Demand*. New York: Elsevier Science.

Hensher, D. A. and F. L. Mannering (1994). Hazard-Based Duration Models and Their Application to Transport Analysis. *Transport Reviews*, Vol. 14, Issue 1, pp. 63-82.

Hettmansperger, T. P. and H. Thomas (2000). Almost Nonparametric Inference for Repeated Measures in Mixture Models. *Journal of the Royal Statistical Society*, Series B, Vol. 62, Issue 4, pp. 811-825.

Horowitz, J (1980). Accuracy of the Multinomial Logit Model as an Approximation to the Multinomial Probit Model of Travel Demand. *Transportation Research*, Part B, Vol. 14, Issue 4, pp. 331-341.

Horowitz, J (1981a). Sampling, Specification and Data Errors in Probabilistic Discrete Choice Models. In D.A. Hensher and L.W. Johnson (*eds.*) *Applied Discrete Choice Modelling.* New York: John Wiley & Sons, pp. 417-435.

Horowitz, J. (1981b). Identification and Diagnosis of Specification Errors in the Multinomial Logit Model. *Transportation Research*, Part B, Vol. 15, Issue 5, pp. 345-360.

Horowitz, J. L. (1997) Accounting for Response Bias: Introduction. In T. F. Golob *et al.* (*eds.*) *Panels for Transportation Planning: Methods and Applications*. Boston: Kluwer Academic Publishers.

Horowitz, J. L. and Manski, C.F. (1998). Censoring of Outcomes and Regressors Due to Survey Non-response. *Econometrica*, Vol. 84, Issue 1, pp. 37-58.

Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge, England: Cambridge University Press.

Imbens, G. (1992). An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica*, Vol. 60, Issue 5, pp. 1187-1214.

Interagency Household Survey Nonresponse Group (1999). Household Nonresponse: What We Have Learned and a Framework for Further Work. *Statistical Policy Working Paper*, Vol. 28, pp. 153-180.

James, F., J. Hoogland and R. Kleiss (1997). Multidimensional Sampling for Simulation and Integration: Measures, Discrepancies, and Quasi-Random Numbers. *Computer Physics Communications*, Vol. 99, pp. 180-220.

Johnson, N. and S. Kotz (1970*). Distributions in Statistics: Continuous Univariate Distributions*, New York: John Wiley.

Kalbfleisch, J., and R. Prentice (1980). *The Statistics Analysis of Failure Time Data.* New York: Wiley.

Kasprzyk, D., *et al.* (*eds.*) (1989). *Panel Surveys*. New York: John Wiley & Sons.

Kiefer, N. M. (1988). Economic Duration Data and Hazard Functions. *Journal of Economic Literature*, Vol. 26, Issue 2, pp. 646-679.

Kitamura, R. and P.H.L. Bovy (1987). Analysis of attrition biases and trip reporting errors for panel data. *Transportation Research*, Part A, Vol. 21, Issue 4-5, pp. 287-302.

Kitamura, R. (1990). Panel analysis in transportation planning: an overview. *Transportation Research*, Part A, Vol. 24, Issue 6, pp. 401-415.

Kitamura, R. and P.H.L. Bovy (1990). Heterogeneity and state dependence in household car ownership: A panel analysis using ordered-response probit models with error components. In M. Koshi (*ed.*) *Transportation and Traffic Theory*. New York: Elsevier Science, pp. 477-496.

Kong, A., J. S. Jun and W. H. Wong (1994). Sequential Imputations and Bayesian Missing Data Problems. *Journal of the American Statistical Association*, Vol. 89, Issue 425, pp. 278-288.

Korimilli, M. S., R. M. Pendyala and E. Murakami (1998). Metaanalysis of Travel Survey Methods. *Transportation Research Record*, Vol. 1625, pp. 72-78.

Kurth, D. L., J. L. Coil and M. J. Brown (2001) An Assessment of Quick-Refusal and No-Contact Nonresponse in Household Travel Surveys. Paper presented at the 80th annual meeting of the Transportation Research Board.

Laaksonen, S. (1999). How to Find the Best Imputation Technique? Tests with Three Methods. Draft for the International Conference on Nonresponse. [Online] Available: http://www.jpsm.umd.edu/icsn/papers/Laaksonen.htm. [May 30th, 2002].

Lancaster, T. (1979). Econometric Methods for the Duration of Unemployment. *Econmetrica*, Vol. 47, Issue 4, pp. 939-956.

Lancaster, T. (1985). Generalized Residuals and Heterogeneous Duration Models with Application to the Weibull Model. *Journal of Econometrics*, Vol. 28, pp. 155-169.

Lancaster, T. and G. Imbens (1990). Choice-based sampling of dynamic populations. In J. Hartog, *et al.* (*eds.*) *Panel Data and Labor Market Studies*. New York: Elsevier Science, pp. 21-43.

Lee, L. (1992). On efficiency of methods of simulated moments and maximum simulated likelihood estimation of discrete response models. *Econometric Theory*, Vol. 8, pp. 518-552.

Lillard, L. A. and R. J. Willis (1978). Dynamic Aspects of Earning Mobility. *Econometrica*, Vol. 46, Issue 5, pp. 985-1012.

Lillard, L. A. (1993). Simultaneous Equations for Hazards: Marriage Duration and Fertility Timing. *Journal of Econometrics*, Vol. 56, pp. 189-217.

Lillard, L. A. and C. W. A. Panis (1998). Panel Attrition from the Panel Study of Income Dynamics: Household Income, Marital Status, and Mortality. *The Journal of Human Resources*, Vol. 33, Issue 2, pp. 437-457.

Little, R. J. A. (1982). Models for Nonresponse in Sample Surveys. *Journal of the American Statistical Association*, Vol. 77, Issue 378, pp. 237-250.

Little, R. J. A. and D.B. Rubin (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Loosveldt, G., J. Pickery and J. Billiet (1999). Item Non-response as a Predictor of Unit Non-response in a Panel Survey (first draft). Paper presented at the International Conference on Survey Non-response, Portland, OR.

Longford, N. T. (2000). Handling Missing Data in Diaries of Alcohol Consumption. *Journal of the Royal Statistical Society*, Series A, Vol. 163, Issue 3, pp. 381-402.

Luce, R. D. (1959). *Individual Choice Behavior*. New York: Wiley.

Ma, J. and K. G. Goulias (1996). Sample Weights for Puget Sound Transportation Panel Using Stratification Anchors in Public Use Microdata Sample and Probabilistic Models for Self-Selection. *Transportation Research Record*, Vol. 1551, pp. 36-44.

Madow, W. G., I. Olkin and D.B. Rubin (*eds.*) (1983). *Incomplete Data in Sample Surveys*, Vol. 2, Vol. 3. New York: Academic Press.

Mannering, F., E. Murakami and S-G. Kim (1994). Temporal stability of travelers' activity choice and home-stay duration: some empirical evidence. *Transportation*, Vol. 21, pp. 371-392.

Manski, C. F. and S. Lerman (1977). The estimation of choice probabilities from choice based samples. *Econometrica*, Vol. 45, Issue 3, pp. 1977-1988.

Manski, C. F. (1994). The selection problem. In C. Sims (*ed.*) *Advances in Econometrics: Sixth World Congress*. Cambridge, England: Cambridge University Press.

Marchak, J. (1960). Binary Choice Constraints on Random Utility Indicators. In K. Arrow (*ed.*) *Stanford Symposium on Mathematical Methods in the Social Sciences*. Stanford University Press.

McCarthy, J. S. and D. G. Beckler (1999). An Analysis of the Relationship between Survey Burden and Nonresponse: If We Bother Them More, Are They Less Cooperative? Draft prepared for the International Conference on Survey Non-response, Portland, OR. [Online] Available: http://www.jpsm.umd.edu/icsn/papers/McCarthyBeckler.htm [May 30th, 2001].

McFadden, D. (1989). A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration. *Econometrica*, Vol. 57, Issue 5, pp. 995-1026.

McFadden, D. (1996). Lectures on Simulation-Assisted Statistical Inference. [Online] Available: http://emlab.berkeley.edu/wp/mcfadden1296/mcfadden.ps [June 29th, 2002].

McFadden, D. and K. Train (2000). Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, forthcoming. [Online] Available: http://emlab.berkeley.edu/wp/mcfadden0500/mcfadden0500.pdf [June 29th, 2002].

McFadden, D. (2000). Disaggregate Behavioral Travel Demand's RUM Side: A 30-Year Retrospective. Paper presented at the International Association for Travel Behavior Conference, Gold Coast, Australia. [Online] Available: http://emlab.berkeley.edu/wp/mcfadden0300.pdf [June 29th, 2002].

Meurs, H and G. Ridder (1997). Attrition and Response Effects in the Dutch Mobility Panel. In T. F Golob *et al.* (*eds*.) *Panels for Transportation Planning: Methods and Applications.* Boston: Kluwer Academic Publishers.

Meyer, B. D. (1987). *Semiparametric Estimation of Duration Models.* Ph.D. Thesis, MIT, Cambridge, Massachusetts.

Morokoff, W. J. and R. E. Caflisch (1994). Quasi-Random Sequences and Their Discrepancies. *SIAM Journal on Scientific Computing*, Vol. 15, Issue 6, pp. 1251-1279.

Moustaki, I. and M. Knott (2000). Weighting for Item Non-Response in Attitude Scales by Using Latent Variable Models with Covariates. *Journal of the Royal Statistical Society*, Series A, Vol. 163, Issue 3, pp. 445-459.

Murakami, E. and W.T. Watterson (1990). Developing a Household Travel Panel Survey for the Puget Sound Region. *Transportation Research Record*, Vol. 1285, pp. 40-46.

Murakami, E. and W. T. Watterson (1991). Attrition and Replacement Issues in the Puget Sound Transportation Panel. Paper presented at the 70th Annual Meeting of the Transportation Research Board, Washington, D.C.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Nijman, T. and M. Verbeek (1992). Nonresponse in Panel Data: The Impact on Estimates of a Life Cycle Consumption Function. *Journal of Applied Econometrics*, Vol. 7, Issue 3, pp. 243-257.

Nobile, A., Bhat, C. R. and E. I. Pas (1993). A Random Effects Multinomial Probit Model of Car Ownership Choice. In C. Gatsonis, *et al*. (*eds.*) *Case Studies in Bayesian Statistics*, Vol. III, pp. 419-434.

Pendyala, R. M., K. G. Goulias, R. Kitamura and E. Murakami (1992). Development of weights for a choice-based panel sample with attrition. *Transportation Research*, Part A, Vol. 27, Issue 6, pp. 477-492.

Pendyala, R. M. and R. Kitamura (1997). Weighting methods for attrition in choice-based panels. In T. F. Golob *et al.* (*eds*.) *Panels for Transportation*

Planning: Methods and Applications. Boston: Kluwer Academic Publishers, pp. 233-257.

Press, W.H., S.A. Teukolsky and M. Nerlove (1992). *Numerical Recipes in C: The Art of Scientific Computing.* Massachusetts: Cambridge University Press.

Puget Sound Regional Council (1997). Puget Sound Transportation Panel Survey 1989-1996: Documentation and Survey Instruments.

Raj, D. (1968). *Sampling Theory*. New York: McGraw-Hill.

Raimond, T. and D. A. Hensher (1997). A review of Empirical Studies and Applications. In T. F. Golob *et al.* (*eds.*) *Panels for Transportation Planning: Methods and Applications.* Boston: Kluwer Academic Publishers, pp. 15-72.

Richardson, A. J., E. S. Ampt and A. H. Meyburg (1995) *Survey Methods for Transport Planning*. Melbourne, Australia: Eucalyptus Press.

Richardson, A. J. (2000). Behavioural Mechanisms of Non-Response in Mailback Travel Surveys. Paper presented at the 79th Annual Meeting of the Transportation Research Board, Washington, D.C.

Ridder, G. (1990). Attrition in multi-wave panel data. In J. Hartog, et al. (*eds.*) *Panel Data and Labor Market Studies*. New York: Elsevier Science, pp. 45-69.

Ridder, G. (1992). An Empirical Evaluation of Some Models for Non-Random Attrition in Panel Data. *Structural Change and Economic Dynamics*, Vol. 3, Issue 2, pp. 337-355.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, Vol. 63, Issue 3, pp. 581-592.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, Vol. 4, pp. 87-94.

Rubinstein, R. Y. (1986). *Monte Carlo Optimization, Simulation and Sensitivity of Queueing networks.* New York: John Wiley & Sons.

Schenker, N. and J. M. G. Taylor (1996). Partially Parametric Techniques for Multiple Imputation. *Computational Statistics and Data Analysis*, Vol. 22, pp. 425-446.

Sen, A. *et al.* (1995). Household Travel Survey Nonresponse Estimates: The Chicago Experience. *Transportation Research Record*, Vol. 1493, pp. 170-177.

Snellen, D., *et al.* (2001). Spatial Variability in Response Rates and Data Quality of a Designated Days-Leave Behind-Full Activity Diary. Paper presented at the 80th annual meeting of the Transportation Research Board, Washington, D. C.

Srinivasan, K. K. and H. S. Mahmassani (2000). Analyzing Heterogeneity and Unobserved Structural Effects in Route-Switching Behavior under ATIS: A Dynamic Kernel Logit (DKL) Formulation. Paper presented at the 79th annual meeting of the Transportation Research Board, Washington, D. C.

Steegh, C. (1981). Trends in Nonresponse Rates. *Public Opinion Quarterly*, Vol. 40, pp. 45-57.

Taris, T. W. (1996). Modeling Nonresponse in Multiwave Panel Studies Using Discrete-Time Markov Models. *Quality and Quantity*, Vol. 30, pp. 189-203.

Tezuka, S. (1995). *Uniform Random Numbers: Theory and Practice*. Boston: Kluwer Academic Publishers.

Thakuriah, P. *et al.* (1996). Nonresponse and Urban Travel Models. *Transportation Research Record*, Vol. 1551, pp. 82-87.

Tourangeau, R., M. Zimowski and R. Ghadialy (1997). An Introduction to Panel Surveys in Transportation Studies. Report prepared by National Opinion Research Center for FHWA. [Online] Available: http://tmip.fhwa.dot.gov/clearinghouse/docs/surveys/panel_surveys/ [March 2nd, 2002].

Train, K. (1999). Halton Sequences for Mixed Logit, working paper, University of California at Berkeley.

Train, K. (2002). *Discrete Choice Methods with Simulation*. [Online] Available: http://elsa.berkeley.edu/books/train1201.pdf [June 29th, 2002].

Vella, F. (1998). Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources*, Vol. 33, Issue 1, pp. 127-169.

Verbeek, M. and T. Nijman (1992). Testing for Selectivity Bias in Panel Data Models. *International Economic Review*, Vol. 33, Issue 3, pp. 681-703.

Vilcassim, N. J. and D. C. Jain (1991). Modeling Purchase-Timing and Brand-Switching Behavior Incorporating Explanatory Variables and Unobserved Heterogeneity. *Journal of Marketing Research*, Val. 18, pp. 29-41.

Wilson, E. (1999). Research Practice in Business Marketing: A Comment on Response Rate and Response Bias. *Industrial Marketing Management*, Vol. 28, pp. 257-260.

Xu, R. and J. O'Quigley (2000). Proportional Hazards Estimate of the Conditional Survival Function. *Journal of the Royal Statistical Society*, Series B, Vol. 62, Issue 4, pp. 667-680.

Yee, J. L. and D. A. Niemeier (2000). Analysis of Activity Duration Using the Puget Sound Transportation Panel. *Transportation Research*, Part A, Vol. 34, pp. 607-624.

Yu, J., and H. Cooper (1983). A Quantitative Review of Research Design Effects on Response Rates to Questionnaires. *Journal of Marketing Research*, Vol. 20, pp. 36-44.

Zimowski, M., *et al.* (1997) Nonresponse in Household Travel Surveys. Report prepared by NORC for FHWA. [Online] Available: http://tmip.fhwa.dot.gov/clearinghouse/docs/surveys/nonresponse [August 20th, 2001].

Vita

Huimin Zhao, the daughter of You-Ru Zhao and Hua-Ying Ding, were born in Nanchang, Jiangxi Province, People's Republic of China on December 3rd, 1971. After graduated from Nanchang No. 2 Middle School in 1989, Huimin entered Tongji University in Shanghai and received a Bachelor of Science from the Department of Road and Traffic Engineering in 1993, where she worked as an administrative staff for undergraduate affairs and an assistant engineer for road design and transportation planning projects. Before she joined the transportation program in the University of Texas at Austin, she enrolled at Northeastern University in Boston, Massachusetts and received a Maser of Science in Civil Engineering.

This dissertation was typed by the author.